T

Tableau Économique

Loïc Charles

Abstract

The *Tableau économique* is an important landmark in the history of economics and the earliest attempt to provide a calculable model that can help government in policy making. François Quesnay, who invented the *Tableau* in 1758, provided several versions of it, both in equilibrium and in disequilibrium, to account for the many cases and policies he discussed in his economic writings.

Keywords

Hoarding; Input–output analysis; Leontief, W; Luxury consumption; Marx, K. H; Mathematics and economics; Mirabeau, V. R; Physiocracy; Productive and unproductive consumption; Quesnay, F; Tableau économique

JEL Classification

B

The *Tableau économique* is generally considered the first economic model in its own right ever conceived. François Quesnay wrote the first version (or 'edition') in December 1758 and sent it to his most prominent disciple, the marquis de Mirabeau, a few days later. Fortunately, the letter and a manuscript copy of this first version have survived (INED 2005, pp. 391–403). Over the next 10 years, Quesnay produced numerous versions of his *Tableau économique.* Some of these (the second and third versions) he printed privately in Versailles. The others were published in books he co-wrote with Mirabeau, in articles he published in physiocratic periodicals and in *Physiocratie,* a collection of his own essays edited and published in 1767 by Pierre Samuel Du Pont de Nemours, one of his young disciples. There are also several manuscript versions in Quesnay's own hand in Mirabeau's papers at the French National Archives (INED 2005). However, Quesnay's deep interest in this visual device did not appeal to his contemporaries (Van den Berg 2002). Even his disciples, who like Mirabeau (1775, pp. 203–4) praised him, calling him a new 'Socrates' and the 'Confucius of Europe', felt uneasy with the *Tableau,* and, after Quesnay died in 1774, the *Tableau* was ignored in the history of economics until Karl Marx rediscovered it in 1862 (Gherke and Kurz 1995). It was only with the transformation of economics into a mathematical and diagrammatic science that the *Tableau économique* gained a more important place in economics. At the same time, the *Tableau* has become an object of controversy for both economists and historians. Of the several problems that have been raised, here we have selected the most significant ones. What is the *Tableau*

Macmillan Publishers Ltd (ed.), *The New Palgrave Dictionary of Economics*,
https://doi.org/10.1057/978-1-349-95189-5

économique? What are its origins? What was its purpose for Quesnay? And how to interpret it? We will conclude this article by assessing the place of Quesnay's *Tableau économique* in the history of economics.

Our understanding of the nature of the *Tableau économique* and its place in Quesnay's economics has deepened since the rediscovery of the 'third version (or edition)' of the *Tableau* by Marguerite Kuczynski in 1965 and the publication in 2005 of several manuscript versions of the *Tableau* from Quesnay's own hand (Kuczynski and Meek 1972; INED 2005). It is now clear that the *Tableau économique* per se is only the figure and as such 'just a small part of Quesnay's model' (Pressman 1994, p. 5). Hence, the 'first edition' consists of three parts: the *Tableau* itself, Quesnay's marginal commentary explaining its working, and the 22 'remarks on the variations in the distribution of the annual revenue of a nation', an expansion of the 14 maxims from the article 'Grains' published in 1757 and which can be best described as a list mixing hypotheses and policy recommendations (see INED 2005, pp. 198–212). The second and third versions of the *Tableau,* compiled in 1759, are based on the same structure, despite formal changes. The 'Remarks' (23 in the second and 24 in the third version) are presented as an 'extract of the royal economic memoirs of M. de Sully'. The commentary, still in the margins in the second version, is transformed into an 'Explanation of the *Tableau économique'* in the third version. In the version of the *Tableau* found in the 'Analyse de la formule arithmétique du Tableau économique' published in *Physiocratie* (1767), only the *Tableau* and its explanation (now called 'analysis') remained in the same text; the maxims are now expanded to 30, with more numerous and longer notes, and presented in an independent text, 'General maxims for the economic government of an agricultural kingdom'.

The origins of the *Tableau* have been seen as a problem since the end of the nineteenth century, when it was first remarked that there was a likeness between the zigzag diagram of the first three versions and the process of blood circulation. This hypothesis was developed by Foley (1974) and further refined by Christensen (1994). However, other scholars have criticized this theory and pointed out there was no direct evidence in Quesnay's writings to support it (Rieter 1983; Eltis 1984, pp. 20–1). Since 1990, several studies have established that the *Tableau* can be seen as a sort of technological device designed to calculate economic quantities as well as display economic laws. Its features derived from different areas of knowledge, covering, for example, art works, machines and early modern arithmetic, which Quesnay brought together to produce a unique concept (Charles 2000, 2003, 2004; Rieter 1990; Wise 1990). Quesnay invented the *Tableau économique* as a visual calculating machine or, as he and Mirabeau put it in the introduction to *Philosophie rurale,* an 'arithmetic rule' and a 'formula of calculation' (Mirabeau and Quesnay 1763, pp. xix, xxiii).

This leads us to the more general problem of the economic interpretation of the *Tableau,* which has captivated economists since Karl Marx. There have been several articles and books providing interpretations, but no general consensus has emerged and important theoretical aspects of the *Tableau* have been controversial ever since Marx's rediscovery of the *Tableau*. One issue that had been discussed at length by commentators is whether the different versions, in particular the zigzag and the late version from 'Analyse de la formule arithmétique', are consistent. The most prominent scholar of physiocracy, R.L. Meek (1962), thought they were, as did several commentators writing under his influence (Barna 1975; Eagly 1969; Pressman 1994). Indeed, the various versions of the *Tableau économique* are based on a single microeconomic unit: a farm made of one plough, four horses and 120 acres of land which is considered as the best technique of production available to agricultural entrepreneurs (Eltis 1984, pp. 3–13; Herlitz 1996, p. 17; Cartelier 1998, p. 249). There are also features common to all versions of the *Tableau*. First, all the *Tableaux* are divided into three columns representing the three classes of citizen relevant to Quesnay's economic representation of society: the productive class (the *classe productive* composed of farmers and professions linked to the primary sector and the marketing of its products), the landowning class (the *classe des propriétaires* which includes landowners, the state and the

clergy) and the unproductive class (the *classe stérile*, corresponding roughly to those working in commerce and industry). Second, the *Tableaux* feature a general interdependence between the three classes in the form of rows linking them: these rows signal the mutual expenditures between the classes that take place in the economic system.

However, the two main versions of the *Tableaux* presented significant differences. First, the *Tableau* in zigzag is an open system that functions as a 'table of expenditure and reproduced produce', and includes a propagation effect that resembles the Keynesian multiplier and accelerator effect (Herlitz 1996, p. 13; Hishiyama 1960, pp. 124–6; Meek 1962, p. 293). Conversely, the mature version of the *Tableau économique* was designed to provide 'a consistent account of social reproduction' as a whole and leave out many details of the process of expenditure and reproduction to concentrate on the general results (Barna 1975; Herlitz 1996). These two main versions of the *Tableau* are also based on different economic models (Cartelier 1982; Herlitz 1961, 1996). In the zigzag version, emphasis is put on the process of expenditure: it is the spending of the landowning class that comes first in the graphical representation, and is seen to initiate the circulation of wealth in the economy (see the two diagrams in Quesnay, François). In Quesnay's economic model, the prominence of the landowning class is made clear by the fact that the equilibrium of the whole system depended on their expenditure. When landowners spend one-half of their revenue on agricultural products and the other on industrial goods, the *Tableau* (and the model) is in equilibrium and simply reproduces itself without changes. When landowners spend more on industrial goods – indulging in *'luxe de décoration'* (excess consumption of luxury goods) – equilibrium is disrupted and the *Tableau* as well as the economy are in decline. Conversely, when they spend more on agricultural products – indulging in *'faste de subsistence'* (increased consumption of subsistence goods) – the economy grows and produces a larger economic surplus.

In the latest version of the *Tableau*, emphasis has switched to the expenditure of the productive class in annual advances (invested in production) and in rent payments to the landowners. In this version, the prominent role goes to the farmers and their advances (which do not figure in the zigzag version), which initiate the process of reproduction of wealth. Conversely, the landowners' expenditure now appears contingent and the landowning class unnecessary for the functioning of the system: the fact that it is the landowners who seize the disposable surplus in the form of rent is arbitrary and linked to the historical context of the society depicted by Quesnay (*Ancien Régime* France), but has no economic justification (see Cartelier 1982). The coexistence of two alternative versions of the *Tableau* corresponded to two policy issues underlined by Quesnay in his economic maxims: (*a*) the necessity of productive advances and (*b*) both the moral and the economic imperative of virtuous spending on the part of the landowning class (no excess luxury consumption). At the same time, it signals problems that Quesnay was unable to resolve satisfactorily. As several commentators have noted, in the latest version of the *Tableau* the landowners' consumption has no bearing on the economic equilibrium, hence there is no need for the hypothesis that landowners spend half their money on agricultural products and the other half on industrial goods; it is enough that they spend all their revenue and do not hoard (Barna 1975; Bilingsoy 1994; Cartelier 1982; Negishi 1989).

The place of the *Tableau économique* in the history of economics has increased in importance dramatically from the 1940s onwards. It has been interpreted either as a forerunner of neoclassical general equilibrium (Samuelson 1982; Schumpeter 1954), as a Leontief input–output system and more generally a linear system (see Phillips 1955; Maitel 1972; Barna 1975; Bilingsoy 1994), while Marxist authors have interpreted the *Tableau* as a rationalization of *Ancien Régime* society (Fox-Genovese 1976; Gleicher 1982). The first interpretation, notwithstanding the authority of Schumpeter (and Samuelson!) holds only at the more general level since there is no trace of marginal analysis in Quesnay's economic model. There is much more ground to link the *Tableau économique* to input–output analysis (for a different opinion, see Pressman 1994, ch. 5). Indeed, Leontief himself

has suggested, if rather cryptically, that Quesnay's work had been an important landmark in the development of his own ideas (Leontief 1936, p. 105). Moreover, like Leontief, Quesnay was interested in providing both a theoretical and an empirical model of the economy at the same time (Barna 1975; Meek 1962, p. 296). According to the Marxist interpretation, the *Tableau économique* exemplified the contradiction of Quesnay's social thought, and more generally of *Ancien Régime* France, caught between the feudal order, characterized by the prominence of landlords, and the burgeoning capitalist economy, characterized by the role played by the capital investments of farmers. Since the fall of the Berlin Wall in 1989 this interpretation has been on the wane among economists and historians alike.

More recently, the attention of economists interested in Quesnay's economics has turned towards several disequilibrium and underemployment-of-resources equilibrium *Tableaux* (Barna 1976; Charles 2000; Eltis 1984, ch. 2; 1996; Pressman 1994, chs 4, 6 and 7). These are particularly noteworthy since they are concerned with the possibility of growth or decline and are linked to specific policy issues. There is no place here to detail these different figures (more than 50 in total!), but it may be useful to list the different economic cases investigated which give rise to these *Tableaux*. First, Quesnay used the *Tableau* to study the effects of hoarding and of excessive consumption of luxury goods by the landowning class. Second, Quesnay investigated the consequences of an unjust tax system with the case of a tax on the productive sector (in Quesnay's theory tax should be levied on the landowning class) and the case of the *Ancien Régime's* (costly) taxation system (the *fermes générales).* Third, Quesnay discussed the joint cases of low agricultural prices due to impediments in external trade and of the economic effects of the establishment of free export in agricultural products (and the rise in prices it causes). Finally, other *Tableaux* are used to show the consequences of policies encouraging the unproductive industrial sector at the expense of the productive agricultural sector.

All in all, Quesnay's *Tableau économique* is now considered by most economists as an important landmark in the history of economics and the earliest attempt to provide a calculable model that can be used by governments for policymaking.

See Also

- ▶ Classical Growth Model
- ▶ Du Pont de Nemours, Pierre Samuel (1739–1817)
- ▶ Ephémérides du citoyen ou chronique de l'esprit national
- ▶ Physiocracy
- ▶ Quesnay, François (1694–1774)

Bibliography

Barna, T. 1975. Quesnay's *Tableau* in modern guise. *Economic Journal* 85: 485–496.

Barna, T. 1976. Quesnay's model of economic development. *European Economic Review* 8: 315–338.

Bilingsoy, C. 1994. Quesnay's *Tableau économique:* Analytics and policy implications. *Oxford Economic Papers* 46: 519–533.

Cartelier, J. 1982. De l'ambiguïté du Tableau économique. *Cahiers d'économie politique* 9: 39–63.

Cartelier, J. 1998. Quesnay, François, and the Tableau économique. In *The Elgar companion to classical economics*, ed. H. Kurz and N. Salvadori. Cheltenham/Northampton: Edward Elgar.

Charles, L. 2000. From the *Encyclopédie* to the *Tableau économique:* Quesnay on freedom of the grain trade and economic growth. *European Journal of the History of Economic Thought* 7: 1–22.

Charles, L. 2003. The visual history of the Tableau Économique. *European Journal of the History of Economic Thought* 10: 527–550.

Charles, L. 2004. The Tableau Économique as rational recreation. *History of Political Economy* 36: 445–474.

Christensen, P.P. 1994. Fire, motion, and productivity: The proto-energetics of nature and economy in François Quesnay. In *Natural images in economic thought*, ed. P. Mirowski. Cambridge, MA: Cambridge University Press.

Eagly, R. 1969. A physiocratic model of dynamic equilibrium. *Journal of Political Economy* 77: 66–84.

Eltis, W. 1984. *The classical theory of economic growth*. London: Macmillan.

Eltis, W. 1996. The *Grand Tableau* of François Quesnay's economics. *European Journal of the History of Economic Thought* 3: 21–43.

Foley, V. 1974. An origin of the Tableau économique. *History of Political Economy* 5: 121–150.

Fox-Genovese, E. 1976. *The origins of physiocracy: Economic revolution and social order in eighteenth-century France*. Ithaca: Cornell University Press.
Gherke, C., and H. Kurz. 1995. Karl Marx on physiocracy. *European Journal of the History of Economic Thought* 2: 53–90.
Gleicher, D. 1982. The historical bases of physiocracy. *American Journal of Economics and Sociology* 46: 328–360.
Herlitz, L. 1961. The *Tableau économique* and the doctrine of sterility. *Scandinavian Economic History Review* 9: 11–51.
Herlitz, L. 1996. From spending reproduction to circuit flow and equilibrium: The two conceptions of *Tableau économique. European Journal of the History of Economic Thought* 3: 1–20.
Hishiyama, I. 1960. The *Tableau économique* of François Quesnay. *Kyoto University Economic Review* 30: 1–46.
INED (Institut National d'Études Démographiques). 2005. François Quesnay. In *Œuvres économiques complètes et autres textes*, vol. 2, ed. C. Théré, L. Charles, and J.C. Perrot. Paris: INED.
Kuczynski, M., and R.L. Meek. 1972. *Quesnay's Tableau économique*. London/New York: Macmillan and Augustus M. Kelley.
Leontief, W. 1936. Quantitative input and output relations in the economic system of the United States. *Review of Economic Statistics* 18: 105–125.
Maitel, S. 1972. The *Tableau Économique* as a Leontief model: An amendment. *Quarterly Journal of Economics* 86: 504–507.
Meek, R. 1962. *The economics of physiocracy*. London: George Allen & Unwin.
Mirabeau, V.-M. 1775. Eloge funèbre de François Quesnay. *Nouvelles Ephémérides Economiques* 1: 197–216.
Mirabeau, V.-M., and F. Quesnay. 1763. *Philosophie rurale, ou économie générale et politique de l'agriculture réduite à l'ordre immuable des loix physiques et rurales, qui assurent la prospérité des empires*. Amsterdam: Libraires associés.
Negishi, T. 1989. Expenditure patterns and international trade in Quesnay's *Tableau économique*. In *Developments in Japanese economics*, ed. R. Sato and T. Negishi. Tokyo: Academic.
Phillips, A. 1955. The *Tableau économique* as a simple Leontif model. *Quarterly Journal of Economics* 69: 134–144.
Pressman, S. 1994. *Quesnay's Tableau économique. A critique and reassessment*. New York: Augustus M. Kelley.
Quesnay, F. 1767–1768. In *Physiocratie ou constitution naturelle du gouvernement le plus avantageux au genre humain*, ed. P.-S. Du Pont [de Nemours]. Merlin: Leyde and Paris.
Rieter, H. 1983. *Zur Reception der Kreislaufanalogie in der Witschaftswissenchaft. Studien zur Entwicklung der ökonomischen theorie III*. Berlin: Duncker-Humblot.
Rieter, H. 1990. *Quesnays Tableau Économique als Uhren-Analogie. Studien zur Entwicklung der ökonomischen theorie IX*. Berlin: Duncker-Humblot.
Samuelson, P.A. 1982. Quesnay's *Tableau économique* as a theorist would formulate it today. In *Classical and Marxian political economy*, ed. I. Bradley and M. Howard. London: Macmillan.
Schumpeter, J.A. 1954. *History of economic analysis*. London: Allen & Unwin.
Van den Berg, R. 2002. *Contemporary responses to the* Tableau économique. In *Is there progress in economics? knowledge, truth and the history of economic thought*, ed. S. Boehm et al. Cheltenham: Edward Elgar.
Wise, N.M. 1990. Mediations. Enlightenment balancing acts, or the technologies of rationalism. In *World changes: Thomas Kuhn and the nature of science*, ed. P. Horwich. Cambridge, MA: Harvard University Press.

Taking (Eminent Domain)

Perry Shapiro

Abstract

Compensation for a public taking of private property (typically land) can affect both government's and landowners' decisions, and compensation rules affect both real and perceived fairness. Arguably, no compensation is efficient; landowners will account for probable loss when making their investment decisions, since such decisions are determined by events outside the discretion of either the landowners or the government (though if the taking decision is made to benefit government or landowners, zero compensation may be inefficient). But zero compensation is unconstitutional (in the United States and Australia) and inequitable. A trade-off between efficiency and equity is therefore normally unavoidable.

Keywords

Compensation calculus; Efficiency–equity trade-off; Epstein, R; Fairness; Just compensation; Land use; Public property; Rent seeking; Taking; Fifth Amendment (US Constitution)

JEL Classification
H0; K0; K19; H41; D63; D72

The Fifth Amendment of the US Constitution ends with 'nor shall private property be taken for public use without just compensation'. Similarly, the Australian Constitution, in section 51(xxxi), allows '[t]he acquisition of property on just terms . . .'. Both constitutions give government the power to condemn property for a public purpose and both specify the requirement for compensation. However, they leave unanswered the formula for computing 'just compensation'. Some dispute that government should be permitted to force the sale of private property through condemnation, but it is widely held that many functions of government, particularly those that require the assemblage of contiguous plots of land, would be impractically complicated without the power of eminent domain. For most, the question is not whether the government should have the right to compulsory acquisition but, rather, what compensation should be paid to the private landowners.

There are two distinct issues when considering compensation for a public taking of private property. The first is that compensation rules can affect the decisions of both government and landowners. The rate at which landowners must be paid for their condemned property may affect the public's choice of how much land to condemn. In addition, the knowledge of potential compensation can influence landowners' choice of property improvements. If compensation is based on the market value of land, landowners will overimprove their property because they do not consider the possibility of a taking (moral hazard) (Blume et al. 1984); and if improvements increase compensation landowners are inclined to overinvest in their property to favorably affect their settlement with the government (rent seeking) (Fischel, 1995 p. 296).

The second issue is that compensation rules affect both real and perceived fairness, and it is likely that equity is what just compensation is about. In the US Supreme Court 1960 decision on *Armstrong* v. *United States,* the majority opinion was that '[t]he Fifth Amendment guarantee . . . [is] designed to bar Government from forcing some people alone to bear public burdens which, all fairness and justice, should be borne by the public as a whole.' A large group of people, those whose property is not condemned, benefit at the expense of a much smaller minority who must surrender their property. Even when compensation is equal to the pre-taking market value of property, as is most commonly the case, the owners of condemned property lose relative to those escaping condemnation.

Blume et al. (1984) use an example of land in a river valley to illustrate potential moral hazard loss. With known probability, p, the price of oil will rise to a price sufficiently high that it is in the public interest to dam the river and flood the valley. The structures invested on the land are all lost under the reservoir waters. With known loss probability, p, the efficient level of investment on the land equates the expected marginal product (the product of p and the marginal product of capital) with r, the market return on capital. If the landowners are fully compensated for both their lost land and immovable capital, the probability of dam- caused flooding will not affect the level of investment on the land. Their investment choice will equate the marginal product of invested capital with the market rate of return ($MP = r$). The result is that an inefficient amount of capital will be invested in the river valley. The conclusion is that no compensation is efficient because landowners will correctly account for probable loss when making their investment decision.

The recommendation for no compensation is based on the presumption that, while uncertain, the decisions whether or not to condemn land and, if so, how much land are determined by events outside the discretion of either the landowners or the government. However, if the taking decision is made to benefit either the government or the landowners, zero compensation induces inefficient decisions (Fischel and Shapiro, 1988; 1989). For instance, if the government represents the interest of a subset of its citizens (for example, the majority) and does not act to maximize social welfare, the need to make compensatory payments to the owners of condemned land will put a beneficial constraint on the government's propensity to

condemn an inefficiently large amount of land. However, even if the government is venal and self-serving, it is never efficient to compensate landowners for 100 per cent of lost value. For the sake of social efficiency the landowners, even in the face of bureaucratic venality, must account for the probability of a taking, even if the amount is inefficiently chosen.

Efficiency is only part of the public policy story. While it might be efficient for government to take land without compensation, it nonetheless offends our notion of what is fair. It is unlikely that any policy as draconian as the one suggested would be adopted for the physical acquisition of private property. (Public regulations that restrict the use of private property commonly are not thought to require compensation.) Uncompensated condemnation is not consistent with a constitution that spells out both the powers and the limits of government, as does the US Constitution. Uncompensated condemnation appears as a bullying big government strong-arming a small minority of landowners.

Frank Michelman (1967) proposes a compensation calculus incorporating fairness. Michelman includes an explicit and quantifiable measure he labels 'demoralization cost'. Demoralization is the personal (psychological) reaction to a government leviathan that runs roughshod over a landowning minority. It is manifest in two different ways: the outpouring of sympathy for the downtrodden, and a concern that the same can happen to you. Citizens empathize with the taken and, simultaneously, worry about the sanctity of their own property rights.

Whether or not Michelman's calculus is used, it is important for real public policy to balance the potential inefficiencies resulting from compensation with the inequities without it. While in most cases it is impossible to achieve efficiency without sacrificing some degree of fairness, or to achieve a fair outcome without sacrificing efficiency, there are special cases for which this is not true. If somehow the interests of the private landowners and the government can be aligned with social welfare, the investment and taking choice can be equitable as well as efficient.

In his discussion of disproportionate impact, Richard Epstein (1985) argues that, if prospective compensation does not affect investment choices, the interests of the landowners and government are the same if takings require full compensation. The case against full compensation is that it induces inefficient investment choices. Without the resource-use concerns, land prices serve to direct government to make only welfare-increasing decisions about condemnation.

For certain types of public projects – those for which changes in land values reflect all the benefits – compensation equal to the post-taking enhanced land value has favorable efficiency and equity consequences. It is equitable because those who lose their property are rewarded equally to those who are lucky enough to escape condemnation. It is efficient because the possibility of condemnation is independent of individual property improvement.

If the benefits of the public taking are specific with measurable market values, it is possible to devise a compensation scheme that achieves both equity and efficiency. However, the dual goals are unattainable when project benefits are diffuse and immeasurable. With these more common cases, it is necessary to consider the trade-off between equity and efficiency. The Michelman calculus is useful in expressing this trade-off if the underlying quantities (discouragement costs) are truly measurable.

See Also

- ▶ Justice
- ▶ Property Law, Economics and

Bibliography

Blume, L., D. Rubinfeld, and P. Shapiro. 1984. The taking of land: When should compensation be paid? *Quarterly Journal of Economics* 99: 71–92.

Epstein, R. 1985. *Taking: Private property and the power of eminent domain*. Cambridge, MA: Harvard University Press.

Fischel, W. 1995. *Regulatory takings, law economics and the power of eminent domain*. Cambridge, MA: Harvard University Press.

Fischel, W., and P. Shapiro. 1988. Taking, insurance and Michelman: Comments on economic interpretation of

just compensation law. *Journal of Legal Studies* 17: 269–293.
Fischel, W., and P. Shapiro. 1989. A constitutional choice model of compensation for takings. *International Journal of Law and Economics* 9: 115–128.
Michelman, F. 1967. Property, utility and fairness: Comments on the ethical foundations of just compensation law. *Harvard Law Review* 80: 1165–1258.

Tarbell, Ida Minerva (1857–1944)

Alice H. Amsden

Keywords
Fordism; Industrial relations; Marginalism; Scientific management; Tarbell, I. M.; Taylorism

JEL Classifications
B31

In a profession where women have been denied the liberties of expression otherwise permitted to men of lesser quality, Ida M. Tarbell made her mark as an economic journalist. She had a keen sense of ethical issues, but regrettably, her blend of reformism and conservatism was sometimes bewildering.

Born in 1857 in Erie County, Pennsylvania, Tarbell is best known for her *History of the Standard Oil Company* (1904), a two-volume attack on the ruthlessness of the oil monopolies (her father was ruined by them). As a muckraker, Tarbell could be expected to favour state intervention in wage setting, then a hotly debated issue. But ironically, whereas the progenitor of marginal productivity theory, J.B. Clark, edged towards arbitration to reduce labour strife, Tarbell embraced Taylorism, whose logic of work atomization builds on the marginalist principle. From 1912 to 1915 Tarbell toured factories she handpicked to study industrial conditions. Favourably struck with Fordism, she wrote a contemporary equivalent of the 'excellently managed corporation' entitled *New Ideals in Business* (1916). The ideals were scientific management, humanistic labour relations and a belief in the fundamental goodness of entrepreneurs.

In her feminist outpourings, Tarbell is better remembered for the way she lived than by what she wrote. Unusual for a woman at the time, Tarbell moved to Paris after college to study women in the French Revolution, as praised by Woodrow Wilson for her 'common sense' views on the tariff (which she opposed), attended the Paris Peace Conference, corresponded with notables, including Richard T. Ely, interviewed Mussolini, and shunned marriage for a career. The same Tarbell, however, fought against women's suffrage and in *The Business of Being a Woman* (1912) advised members of her sex to stay at home.

Selected Works

1904. *The history of the standard oil company.* New York: McClure, Phillips.
1911. *The tariff in our times.* New York: Macmillan.
1912. *The business of being a woman.* New York: Macmillan.
1915. *The ways of woman*. New York: Macmillan.
1916. *New ideals in business: An account of their practice and their effects upon men and profits.* New York: Macmillan.
1926. *The life of Elbert H. Gray: The story of steel.* New York: D. Appleton.
1936. *The nationalizing of business, 1878–1898.* New York: Macmillan.

Targets and Instruments

Jan Tinbergen

These are two concepts used in the theory of economic policy. Although the subject of economic policy as one of the forms of applied economic theory is as old as economic science itself, the more systematic treatment meant by the phrase 'theory of economic policy' started much more

recently, in close connection with the development of econometrics. Econometrics, as the combination of theory and observation in the area of intersection of economics, statistics and mathematics, introduced the possibility of dealing with economic policy not only qualitatively, but also quantitatively. This enables economists to formulate policy recommendations in the most concrete form conceivable, as they are needed by policy-makers – government, parliament and representatives of social groups. It seems appropriate to consider as the starting document of the theory of economic policy in this sense, Ragnar Frisch's document (1949), written for the United Nations' short-lived Employment Commission, 'A memorandum of price–wage–tax–subsidy policies as instruments in maintaining optimal employment'.

The term 'targets and instruments' refer to economic variables in a special case of a more flexible version of the theory of economic policy, where the more general terms 'aims and means of a policy' are used, which may be qualitative as well as quantitative. An aim may then be the maximization (under possible restrictions) of social welfare, and among the means a reform may appear. Targets are numerical values of variables appearing in a social-welfare function and are supposed a priori to be the values that maximize social welfare. Instruments are quantitative values of means controllable by the policy-maker (cf. Preston and Pagan 1982).

Examples of target variables are employment, current balance-of-payment surplus, current government surplus, income, the rate of inflation, and others. Examples of instrument variables are direct and indirect tax rates, interest, total or specific public expenditures, working hours per week, working weeks per year, age of retirement, wage rates, and so on.

Problems of economic science may be subdivided into two categories: explanatory or analytical problems, and normative or policy problems. The complete mathematical formulation and solution of these problems require the introduction of two more categories of variables, to be called 'exogenous' (or 'data') variables and 'other' (or 'irrelevant') variables. In what follows, the four categories will sometimes be indicated by x (irrelevant), y (target), z (instrument) and u (data) variables. In addition, the mathematical formulation of the two types of problems requires the fulfilment of a number I of equations or relations, numbered i (=1, 2, . . ., I), a number of J of variables x_j, a number K of variables y_k, a number L of variables z_l, and I variables u_i.

The equations will be assumed to be linear, which for small variations is no restriction, but for large variations constitutes a limitation. They will be written:

$$\sum_j a_{ij}x_j + \sum_k b_{ik}y_k + \sum_l c_{il}z_l = u_i, \quad i = 1, 2, \ldots, I \tag{1}$$

Examples of relations are definitions, technical or legal relations, balance equations and behavioural relations such as demand or supply equations for either goods or factors of production (labour types, capital, etc.). The group of I equations (1) describes, in a simplified way, the operation of the economy studied and is called a 'model' of that economy, more particularly when all coefficients a, b and c have been given numerical values obtained from a series of values for all x, y, z and u over some observation period.

The mathematical-statistical (or econometric) methods of estimation will not be discussed here, but the choice of, in particular, the variables x and u will be such as to obtain reliable values of the coefficients. This implies that the coefficients of determination R^2, corrected for the number of degrees of freedom (and then written $\overline{R}^2$) as well as the so-called t-values satisfy certain conditions, usually $\overline{R}$ should be not far below 1 and $ts > 3$, but this ideal is rarely attained.

A problem (and this applies to both types of problem mentioned) can be solved only if the number of unknowns N equals the number of equations' I – this being a necessary but not a sufficient condition. The unknowns for each time unit are, for the explanatory problem, the target variables and the 'other' variables. So we must have:

$$K + J = N \tag{2}$$

For the political problem the unknowns are the instrument and the 'other' variable, and we must have:

$$L + J = N \tag{3}$$

From (2) and (3) we deduce that $K = L$ must apply for the problems to be solvable; that is, *the number of instruments must equal the number of target variables*. Later we will discuss some exceptions to this thesis, but as a general rule our conclusion stands.

For a more concise treatment of our problems it is sometimes preferable to formulate the model in a simplified form by eliminating the irrelevant variables x. This elimination requires J equations and so we are left with $N - J = K = L$ equations, in which only the y, z and u appear. In order to avoid confusion we will now use capital letters for the coefficients:

$$\sum_k B_k y_k + \sum_l C_l z_l = u_i \quad i = 1, 2, \ldots, I \tag{4}$$

For simplicity's sake we will discuss examples where $K = L = 4$. The explanatory problem's solution is obtained by solving (4) for the target variables y:

$$y_k = \sum_l p_{kl} z_l + \sum_i s_{ki} u_i \quad k = 1, 2, \ldots, K \tag{5}$$

The policy problem's solution is found from solving (4) for the instrument variables z:

$$z_l \sum_k q_{lk} y_k + \sum_i t_{li} u_i \quad l = 1, 2, \ldots, L \tag{6}$$

By use of matrix notation these equations might have been written more elegantly, but we shall refrain from doing so. It seems desirable, though, to express verbally the meaning of the coefficients used. Evidently p_{kl} constitutes the change in y_k caused by a unit change in z_l and no change in the other zs or any u. If normalized variables had been used (i.e. variables with a mean equal to 0 and a standard deviation equal to 1, as is customary in sociologists' path analysis) p_{kl} becomes the partial elasticity of y_k with respect of z_l. In both cases p constitutes a measure for the impact of instrument variable z_l on target variable y_k, all other z and all u assumed constant. Inversely and similarly q_{lk} measure the impact of a unit change in target y_k on instrument z_l, all other y and all data u assumed unchanged.

As previously observed the conditions so far mentioned are necessary but not sufficient. Other conditions which must be fulfilled are that equations (4) be neither incompatible nor dependent nor overdetermined. Simple illustrations are the following. If of four unknowns three appear in only two of the equations and the fourth in the other two equations, then the first two equations are overdetermined and the other two are either incompatible or dependent. Overdetermination implies that there is not just one solution but an infinity of them. Incompatibility means that the solution of one of the two equations does not satisfy the other. Dependency of equations means that one equation can be deduced from the other. In that case they do have the same solution, and so the occurrence of one unknown in both equations does no harm, but the solution for the three other unknowns from the two remaining equations is impossible.

An example of a system of equations suffering from non-fulfilment of the conditions just discussed can be found in Tinbergen (1956), Problem 161, using Model 16. As a counterexample without this difficulty, Problem 162 has been added. In these examples the irrelevant variables had not been eliminated first and the complete model containing 17 variables is shown.

Some politicians think that the normal situation is that there is a one-to-one correspondence between particular targets and particular instruments, for example that a tax rate is used to equilibrate the government budget, an exchange rate to equilibrate the balance of payments, and a wage rate to create enough employment. As a rule this is not correct, for such a situation would imply that only one z_l appears in each of the four equations (5) and only one y_k in each of the four equations (6), implying in turn that equations (5) and (6) could be arranged so that only the diagonal elements of the matrices P and Q could be non-zero. The normal situation is that not all elements off the diagonal vanish.

There are however some elements equal to zero in most models. An interesting case is that where

the equations can be ordered so that all elements above the diagonal are nought, or where blocks of elements are equal to zero. Connected with such coefficient matrices is H.A. Simon's concept of the 'order' of an unknown, which in a policy problem corresponds to the instruments and the irrelevant variables. The concept indicates that the unknowns can be solved in a predetermined order only; the one with order 1 depends on one coefficient only, or if a group of unknowns has order 1, they depend on as many coefficients as appear in the group of equations in which these unknowns only appear. A next unknown or group of unknowns depends on the coefficients appearing in the equations containing groups 1 and 2 of the unknowns, and so on. Evidently an ordered system may be organized in a simpler way, because some decision-makers (say, government ministers) can decide quite independently of other decision-makers without deviating from the optimal policy.

Further deviations from the standard case discussed in illustrating equations (5) and (6) will occur if some instrument variables are subject to restrictions, such as the impossibility of negative values or of values less than a previous value. A large number of economic variables (for instance, production and consumption as well as prices) cannot be negative. In today's industrial countries a reduction in nominal wages is almost impossible. If without such a restriction an impossible value of some instrument would be part of the solution, the restriction becomes active; that is, it becomes an equation instead of an inequality. Since the number of unknowns is then less than the number of equations, we have either to add an unknown (an additional instrument) or to omit one equation. A possible example is that a foreign loan may be introduced as an instrument in order to keep the balance of payments in equilibrium.

The introduction of several restrictions (non-negativity of several unknowns) may leave us, after using all the equations to eliminate unknowns, with, say, two unknowns and three restrictions, still permitting any point within a triangle. The latter is called the 'admissible' or 'feasible' area and the remainder of the policy problem may be presented as a problem of linear programming. A choice among the points within this feasible area is now possible by adding the condition that some function of the remaining two unknowns be maximized. The substitution of equations by inequalities need not be used only to express the necessity that a variable be non-negative; a production function may also be interpreted as yielding the maximum quantity of product obtainable from given inputs, any deviation from that maximum then representing waste or 'X-inefficiency'.

Finally, a few words may be said about the use of target and instrument variables in interactive planning (J.A. Hartog, P. Nijkamp, J. Spronk). Frisch and his school built their policy-planning on a social-welfare function obtained by interviewing policy-makers on a universe of local trade-off rates between the variables that determined, in their opinion, the population's level of satisfaction. If n such variables are thought to exist, hypersurfaces of n dimensions would be the hypersurfaces of constant satisfaction. The interaction planning school doubts whether the average policy-maker is able to describe such hypersurfaces. The method they propose is that the policy-planner starts with a given situation and subsequently shows the policy-maker what change in the targets is obtained by an assumed first set of changes in instruments, asking him whether that change in the targets constitutes an improvement. The policy-maker may propose a further change in the instruments and the planner will inform him on the consequences for the targets. Thus, step by step, in this dialogue, planner and policy-maker will approach a situation which does not admit any improvement as a consequence of changes in instruments to be proposed by the policy-maker. In this dialogue the policy-maker will have to compare a limited number of sets of instruments and targets, presumably a much lower number than was needed by the interview method (cf. also Hughes Hallett and Rees 1983).

See Also

▶ Control and Coordination of Economic Activity

Bibliography

Frisch, R. 1949. *A memorandum of price–wage–tax–subsidy policies as instruments in maintaining optimal employment*. United Nations Employment Commission. E/CN.1/Sub 2/13, April.

Hughes Hallett, A., and H. Rees. 1983. *Quantitative economic policies and interactive planning*. Cambridge: Cambridge University Press.

Preston, A.J., and A.R. Pagan. 1982. *The theory of economic policy, statics and dynamics*. Cambridge: Cambridge University Press.

Tinbergen, J. 1956. *Economic policy: Principles and design*. Amsterdam: North-Holland.

Tariff Versus Quota

Arvind Panagariya

Abstract

Bhagwati J. On the equivalence of tariffs and quotas. In: Baldwin RE, Bhagwati J, Caves RE, Johnson HG (eds) Trade, growth and the balance of payments: essays in honor of G. Haberler. Rand McNally, Chicago (1965) first demonstrated that if perfect competition prevails in all markets, a tariff and import quota are equivalent in the sense that an explicit tariff reproduces an import level that, if set alternatively as a quota, produces an implicit tariff equal to the explicit tariff, and vice versa. This equivalence breaks down, for example, if the domestic production is monopolized. In this case, replacing an explicit tariff by an import quota set at the level equal to the imports under the explicit leads to a higher implicit tariff. Many other cases of the breakdown of the equivalence also arise.

Keywords

Directly unproductive profit-seeking; Tariff versus quota; Tariffs; Uncertainty; Voluntary export restraints

JEL Classifications

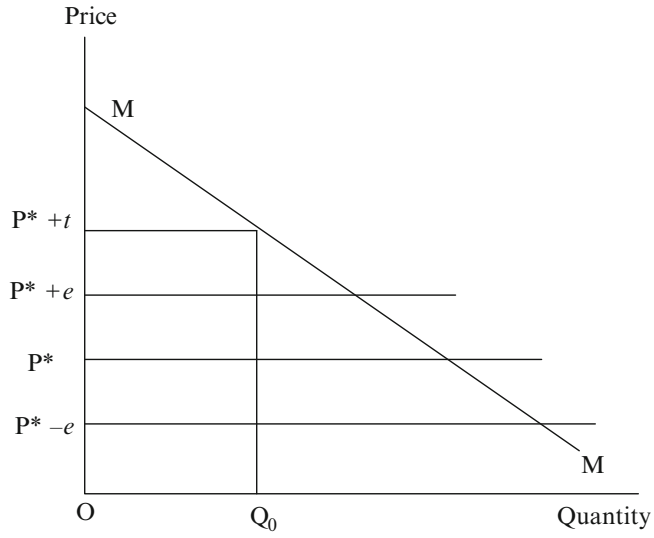
F1

The 'tariffs versus quota' literature was stimulated by the seminal contribution by Bhagwati (1965). Bhagwati defined the two instruments as equivalent if an explicit tariff reproduces an import level that, if set *alternatively* as a quota, produces an implicit tariff equal to the explicit tariff and vice versa.

Equivalence and Its Breakdown

Bhagwati (1965) demonstrated that the tariff–quota equivalence necessarily obtains when perfect competition prevails in all markets. This is shown most simply in the small-country context. By definition, the small country faces a perfectly elastic supply at a given price in the world market. In Fig. 1, DD and SS represent the demand and supply curves in this (small) country and P^* fixed world prices. Under free trade, that country produces Q_0, consumes C_0 and imports Q_0C_0. The imposition of an explicit tariff t per unit raises the internal price in the country to $P = P^* + t$ and the output and consumption move to Q_1 and C_1, respectively. Imports decline to Q_1C_1. The consumer surplus declines by the trapezium formed by the sum of the areas marked b, e, R and f. Of this, area b is recovered by producers as extra surplus and area R by the government as tariff revenue. Areas e and f are lost entirely and called deadweight losses. Area e is lost because the marginal cost of production of Q_0Q_1 exceeds the world price. Area f is lost because the tariff forces the consumers to stop before the marginal benefit at P exceeds the marginal social cost of obtaining the goods at P^*.

If we now replace the tariff by the import Q_1C_1 and auction the quota licences competitively, imports would equal the quota. Subtracting these imports from DD at each price, we obtain $D'D'$ as the demand facing domestic producers. The internal price now obtains at the intersection of SS and $D'D'$. But by construction, this is $P = P^* + t$, with t now representing the implicit tariff. Explicitly, t is now the price of the licence per unit of imports. The total revenue from the auction of the licences equals R. The outcome is identical to that under a tariff in every way.

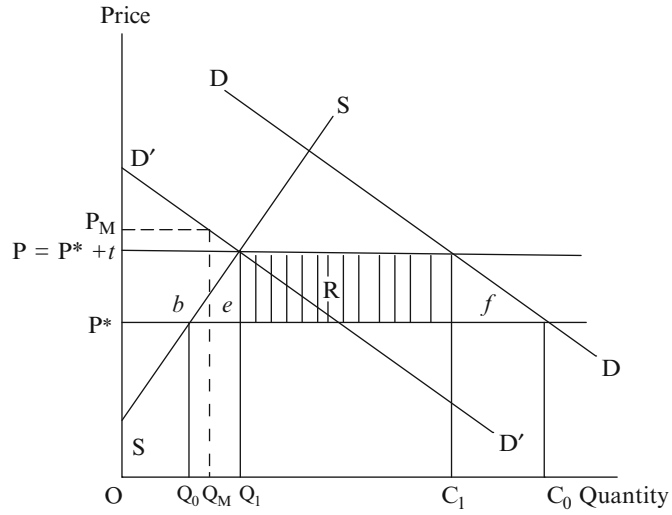


Tariff Versus Quota, Fig. 1 Equivalence under perfect competition and non-equivalence under monopoly

Suppose now that a monopoly producer supplies the domestic output with SS representing his marginal cost curve. Under the tariff t, the monopolist cannot raise the price above $P^* + t$ so that the outcome is no different from under perfect competition. But if we replace the tariff by import quota Q_1C_1, with the quota licences auctioned competitively, the monopolist faces the demand curve D′D′. Associated with D′D′ is a marginal revenue curve (which, for the sake of simplicity, is not shown in Fig. 1) whose intersection with SS gives the monopoly output Q_M. The price the monopolist changes at this quantity is P_M, which is higher than $P^* + t$. The equivalence breaks down. Non-equivalence also obtains if we replace domestic competitive suppliers by an oligopoly rather than a monopoly (Helpman and Krugman 1992). All these results can be generalized to the large-country case. Alternatively, non-equivalence obtains if we assume perfect competition in demand and supply but not the allocation of quota. For example, if the holder of the quota licence is a monopolist, he would maximize the quota rents. The solution in this case may involve leaving some licences, thereby raising the domestic price above $P^* + t$. Retaining perfect competition in all markets, non-equivalence also arises if the quota takes the form of a voluntary export restraint (VER). Under the VER, enforcement of the quota is the responsibility of the exporting country. In this case, the exporting country captures the quota rent, and the welfare loss from the quota to the importing country is larger than under the tariff or direct import quota.

A further case of non-equivalence arises in the presence of uncertainty. This is shown in a simple manner using a construction from Pelcovits (1976). In Fig. 2, suppose the import demand by home country is MM and the world price can be either $P^* + e$ or $P^* - e$, each with a probability of one half. Suppose further that we want to restrict expected imports to OQ_0. Regardless of which world price is realized, an import quota will hit the target exactly with the domestic price given by the height of MM at Q_0 and denoted $P^* + t$. If a tariff is to be used to achieve the same objective, assuming risk-neutral behaviour, we would set the tariff at t. The domestic price in this case could be either $P^* + t + e$ or $P^* + t - e$, each with a probability of one half. The reader can verify that the expected welfare losses under the quota and tariff would be different, which implies non-equivalence.

Panagariya (1980) considers the equivalence of optimal tariff and quota *structures* in a small-country model with two or more imports. He considers a government wishing to restrict the value (at the world prices) of two or more imports to a pre-specified level. If perfect competition



Tariff Versus Quota, Fig. 2 Non-equivalence under uncertainty

prevails in all markets, the optimal policy is either an explicit uniform tariff on the restricted set of goods or import quotas on them at levels that imply a uniform implicit tariff at the same rate. If domestic production of these goods is monopolized, however, optimal tariffs are still uniform, but optimal import quotas are characterized by implicit tariffs that are generally non-uniform.

In two companion papers, Panagariya (1981, 1982) brings out yet another aspect of tariff–quota non-equivalence in the presence of domestic monopoly. He considers a large-country, general-equilibrium model in which domestic industry is monopolized. He shows that in such a model, exogenous changes in quotas and tariffs lead to qualitatively different outcomes. For example, if quota is the instrument, tightening it always improves the terms of trade. But if tariff is the instrument, raising it may lead to deterioration in the terms of trade.

Finally, Rodriguez (1974) and Tower (1975) have independently considered the outcomes when two countries optimally choose trade interventions in a Nash non-cooperative game within a two-good general-equilibrium model. They show that if the countries choose tariff as the instrument, an equilibrium characterized by finite positive trade between them generally exists. But if the countries employ quotas, such equilibrium does not exist.

Welfare Ranking of Tariffs and Quotas

When tariffs and quotas are not equivalent and we use them to target some variable in the economy, a natural question concerns the welfare ranking of the two instruments. For example, if we assume that the domestic production is monopolized but perfect competition prevails everywhere else and the objective is to restrict imports to a specified level at the lowest cost, tariff is superior to quota. This is readily seen in the small-country case in which the tariff forces the monopolist to behave like a competitor whereas the quota allows him to earn positive monopoly rents. Similarly, on the assumption that there is perfect competition on all sides in the product market, if the quota holder is a monopolist, tariff would once again be the superior instrument.

Two additional ranking results are due to Panagariya (1980) and Pelcovits (1976). The former considers the ranking of tariffs and quotas when a small country aims to restrict the value of a subset of imports (at world prices) to a fixed level and these goods are subject to the monopoly distortion. He finds tariffs to be an instrument superior to quotas. Pelcovits considers the welfare ranking in the presence of uncertainty. In the small-country context, on the assumption that the world price is stochastic and the country wishes to constrain the expected imports at a

pre-specified level, he asks whether quotas yield higher expected welfare or tariffs. Using a construction similar to that in Fig. 2, he shows that the answer is ambiguous.

Welfare Outcomes with Pre-existing Tariffs and Quotas

A final question of interest is how the welfare outcomes differ when a parameter is altered in the presence of tariffs versus quotas. The first set of contributions in this category comes from the so-called piecemeal trade reforms literature that asks how welfare changes as we relax one trade barrier at a time. Corden and Falvey (1985) demonstrate that in a small country with many imports, if the country restricts trade by quotas only, the relaxation of any quota necessarily improves welfare. Intuitively, the relaxation of the quota reduces the distortion in that that good has no effect on the distortion in the other goods since their imports face the same quota as before. Therefore, the net effect of the change on welfare is positive. This is not true if imports are restricted by tariffs. A reduction in any one tariff directly improves welfare by expanding the imports of that good. But it may indirectly lower welfare by reducing the imports of substitute goods subject to tariff distortions. If the latter effect dominates, the net effect is a reduction in welfare.

Building on the work of Johnson (1967) and Kemp and Negishi (1970), Eaton and Panagariya (1979, 1982) derive conditions under the presence of tariffs on a subset of imports that can lead to an improvement in the terms of trade or growth in a small open economy to result in a loss of welfare. It is readily shown, however, that if import quotas restrict imports instead, improvement in the terms of trade or growth cannot lead to a decline in welfare. Just as in Corden and Falvey, when quotas are in place, their distortionary effect remains unchanged when the terms of trade improve or growth takes place. Therefore, the direct benefits from improved terms of trade or growth determine the final outcome. In the presence of tariffs, tariff distortion worsens if the improvement in the terms of trade or growth is accompanied by a contraction of imports of one or more tariff-ridden goods.

Finally, Bhagwati and Srinivasan (1982) alternatively consider the effect of directly unproductive profit-seeking (DUP) activities in the presence of tariffs and quotas. They show that in the former case the DUP activity may paradoxically raise welfare if it draws resources out of the import-competing good and therefore leads to an expansion of imports of the tariff-ridden good. This cannot happen in the latter case, however, since the imports of the quota-ridden good cannot rise beyond the fixed import quota.

See Also

- ▶ Tariffs
- ▶ Trade Policy, Political Economy of

Bibliography

Bhagwati, J. 1965. On the equivalence of tariffs and quotas. In *Trade, growth and the balance of payments: Essays in honor of G. Haberler*, ed. R.E. Baldwin, J. Bhagwati, R.E. Caves, and H.G. Johnson. Chicago: Rand McNally.

Bhagwati, J., and T.N. Srinivasan. 1982. The welfare consequences of directly unproductive profit seeking (DUP) lobbying activities: Price vs. quantity distortions. *Journal of International Economics* 13: 15–33.

Corden, W.M., and R.E. Falvey. 1985. Quotas and the second best. *Economics Letters* 18: 67–70.

Eaton, J., and A. Panagariya. 1979. Gains from trade under variable returns to scale, commodity taxation, tariffs and factor market distortions. *Journal of International Economics* 9: 481–501.

Eaton, J., and A. Panagariya. 1982. Growth and welfare in a small, open economy. *Economica* 49: 409–419.

Helpman, E., and P. Krugman. 1992. *Trade policy and market structure*. Cambridge, MA: MIT Press.

Johnson, H.G. 1967. The possibility of income losses from increased efficiency or factor accumulation in the presence of tariffs. *Economic Journal* 77: 151–154.

Kemp, M.C., and T. Negishi. 1970. Variable returns to scale, commodity taxes, factor market distortions and their implications for trade gains. *Swedish Journal of Economics* 72: 1–11.

Panagariya, A. 1980. Import targets and the equivalence of optimal tariff and quota structures. *Canadian Journal of Economics* 13: 711–715.

Panagariya, A. 1981. Quantitative restrictions in international trade under monopoly. *Journal of International Economics* 11: 15–31.
Panagariya, A. 1982. Tariff policy under monopoly in general equilibrium. *International Economic Review* 23: 143–156.
Pelcovits, M. 1976. Quotas versus tariffs. *Journal of International Economics* 6: 363–370.
Rodriguez, C.A. 1974. The non-equivalence of tariffs and quotas under retaliation. *Journal of International Economics* 4: 295–298.
Tower, E. 1975. The optimum quota and retaliation. *Review of Economic Studies* 42: 623–630.

Tariffs

T. Scitovsky

Abstract

Tariffs are taxes levied on goods imported or (less often) exported as they cross a geographical border. They raise revenue but are normally evaluated by reference to their impact on the economy, which usually means the protection they provide to domestic producers and their effect on the terms of trade. Tariffs can exploit a country's monopoly or monopsony position in world markets, but only if that is not already exploited by private firms within the country. An import duty can be used as countervailing power to prevent a country being exploited by a foreign exporter's use of *his* monopoly power.

Keywords

Ad valorem tariffs; Cartels; Devaluation; Effective tariffs; Export tariffs; Extortion tax; Free trade; Great depression; Hamilton, A.; Import substitution; Infant-industry protection; Lerner, A.; List, F.; Monopoly; Monopsony; Protection; Scitovsky, T.; Specific tariffs; Tariffs; Terms of trade; Value added

JEL Classifications

F1

Tariffs are taxes levied on foreign trade: on the importation and, less often, the exportation of goods as they cross the border of a country or other geographical area. Since they are easy to enforce and collect and seem to be (and partly are) paid by foreigners, tariffs have been an important and popular source of government revenue from the earliest times. In early days, the ostensible purpose of tariffs was to pay the government levying them for the protection it afforded to foreign traders on its territory. In modern times, arguments for and against tariffs as well as the determination of their level focus on their impact or supposed impact on the economy.

Tariffs nowadays are paid in money and specified either as so much money per unit of merchandise (specific tariffs) or as a given percentage of its value (ad valorem tariffs). With demand a diminishing function of price, tariffs reduce the quantity of dutiable goods imported or exported; and, with the price elasticity of demand also a diminishing function of price, Government's revenue from the tariff (i.e. the product of its level and the quantity on which it is levied) first increases then diminishes as the tariff rate is raised. Accordingly, there is a rate of tariff that maximizes tariff revenue; but the proper criterion for judging the desirability of tariffs and determining their level is not the amount of revenue they yield but their impact on the whole economy, which is usually discussed under two headings: the protection they provide to domestic producers and their effect on the terms of trade.

Protective Tariffs

Tariffs on imports raise their domestic prices, thereby shifting demand from imports to their domestic substitutes and increasing the profitability of the latter's production. Import duties also lower the purchasing power of income over imports and import substitutes (collectively known as importables) but add to the money incomes of producers of imports substitutes, their employees and suppliers. Accordingly, the tariff-imposing country's real national income may be raised or lowered, depending on whether

the sum of the Government's tariff revenue and the additional incomes generated exceeds or falls short of the loss of purchasing power over importables. That, however, is still only a small part of a full cost-benefit calculation, which must also take into account other costs and benefits of the tariff.

By far the most important among the costs is the danger of retaliation by the foreign countries whose export industries are hurt, or believed to be hurt by the first country's import duties. That cost is especially great when the trade restrictions other countries impose in retaliation to the first country's tariff lead, in their turn, to further retaliations, and so to a general overall reduction in the volume of trade and its gains.

For the impact of import duties is to discourage the imports on which they are levied. It is true that they also stimulate domestic activity and domestic income generation, which, in the long run, may well counteract their restrictive effect on imports, at least to the extent of more or less offsetting the reduction in *overall* imports. But the combined influence of the restrictive short- and expansionary long-run effects of tariffs would have to keep unchanged not only the overall value of total imports but also their structure by country of origin in order to eliminate the economic justification and pressure for retaliation; whereas even the commodity composition of imports would have to remain unchanged in order to eliminate the political pressure for retaliation as well. Needless to say, those conditions necessary to obviate retaliation are not likely to be fulfilled.

The benefits of import duties include increased employment, an improved balance of trade, the enhanced stability of a more diversified economy, the political and economic advantages of greater self-sufficiency, and the increased efficiency of protected industries when their comparative disadvantages are remediable and can be remedied through learning by doing. Some of those advantages, however, are mutually exclusive. Tariffs, for example, that greatly stimulate the domestic economy are unlikely to improve the balance of trade – a fact that was strikingly brought home to many of the developing countries that engaged in import-substitution policies.

Of the benefits listed, by far the most important is the last-mentioned, which is a permanent benefit secured by temporary tariff protection. It has also received the most attention in the professional literature under the name of the infant industry argument. Trade restriction to nurture budding industries was well known and much practised already during the mercantilist period; but after the advent of economic liberalism, the argument in its favour needed to be reasserted. Its best known and most influential statements in modern times are those of Alexander Hamilton and Friedrich List. Hamilton's celebrated 'Report on Manufactures' to the US Congress (1791) had a great influence on American tariff policy, and its prediction of the hoped-for consequences of protection turned out to be a remarkably accurate forecast of the country's subsequent economic development. List's similar argument half a century later (List 1841) had even more influence on both US and German foreign-trade policy.

The US and German protective tariffs of the nineteenth century, however, which seem to have been so successful in promoting those countries' economic development, were very much more moderate than the mid-twentieth-century import barriers behind which India, Pakistan and the Latin-American and other developing countries pursued their not very successful import-substitution policies (Little et al. 1970). That raises the question of what level and structure of protective tariffs are the most conducive to a country's economic development. We cannot answer here that much-debated and highly controversial question; but something must be said about effective tariffs, a statistical tool designed to help the search for an answer.

Effective Tariffs

The height of a tariff levied on imports of a good (also known as the nominal tariff) is not a good measure of the degree of encouragement of its domestic manufacture. For one thing, a manufacturer almost never creates a whole good, only a greater or lesser contribution to it, which is called his value added or effective price; and a given

percentage tariff on imports, which enables domestic manufacturers of its substitutes to raise their prices by a like amount, makes a greater, often very much greater *percentage* addition to their value added, in a proportion that is the inverse of the ratio in which their value added stands to price. For example, if the value added in cloth manufacture is 40% of price, then a 20% nominal tariff on imported cloth enables domestic cloth manufacturers to increase their value added by 50%.

For another thing, tariffs are often levied on final, intermediate and primary goods alike; and an import duty on a primary or intermediate good, while encouraging its domestic production, also discourages the domestic manufacture of all those other goods that use it as an input. An import duty on yarn, for example, *dis*courages domestic cloth manufacture by reducing the value added cloth manufacturers can earn.

The concept of effective tariff (ET) is designed to measure the degree of encouragement provided to given productive activities by the combined effect of the nominal tariffs imposed on their outputs *and* inputs. A simple formula for the effective tariff protection on the manufacture of good j is:

$$ET_j = \frac{t_j}{1 - \sum a_{ij}} - \frac{\sum a_{ij} t_i}{1 - \sum a_{ij}},$$

where t_j is the nominal ad valorem tariff on good j, t_i are the nominal tariffs on its several inputs, and the a_{ij} show the share of the cost of input i in the price of good j at free trade prices. Note that the two terms show the contributions of the two factors discussed in the text, note also that the denominator represents value added as a proportion of price.

The Terms-of-Trade Effect of Tariffs

In contrast to all the attention economists, politicians and the general public have paid to the protection that tariffs provide to domestic industry, the tendency of tariffs to improve the terms on which a country trades its exports for imports, and thereby to increase its share in the gains from international specialization, have been very much neglected. The subject attracted some attention at the end of the last and the beginning of this century, but mainly as a theoretician's intellectual exercise and an economic curiosum.

While protection results from import duties' raising the *domestic* price that *domestic* buyers have to pay for importables, they improve the terms on which the duty-imposing country trades its exports for imports, provided that they lower the *foreign* price that the *foreign* producers of it receive for them. Similarly, an export duty will also improve a country's terms of trade if it raises the *foreign* price that *foreign* buyers of its exports have to pay for them. Accordingly, tariffs improve a country's terms of trade if the foreign supply of its imports or the foreign demand for its exports is less than perfectly elastic; and a given tariff has the greater impact on the terms of trade, the lower are those elasticities (Bickerdike 1906; Kaldor 1940).

The advantage of a tariff that improves the terms of trade can be given two interpretations. First, when a tariff changes the foreign price of imports and/or exports to the foreigners' disadvantage, it causes *them* to pay part of the tariff – a clear and obvious gain for the tariff-imposing country. Secondly, the same gain can also be looked upon as a monopoly or monopsony profit extracted from foreigners by the tariff, which in turn closely resembles the profit margin a monopolist adds on to marginal cost, or a monopsonist subtracts from marginal worth, when he sets his profit-maximizing price. Indeed, when perfect competition among a country's export producers causes them to equate prices to marginal costs and causes its importers to equate the marginal value product of imports to their prices, then export and import duties coincide exactly with a monopolist's and monopsonist's profit margins.

Such a situation resembles a cartel agreement among domestic competitors with respect to their foreign transactions, except that the monopoly or monopsony profits generated accrue to the State in the form of tariff revenue and that the private producers and traders are made worse off than they would be under free trade, because the tariff reduces the volume of their business. From the

point of view of the country's national welfare, however, tariffs can be beneficial, in the sense of increasing the sum of the country's private and public real income, just as monopolistic or monopsonistic pricing can increase the monopolist's or monopsonist's profit. Indeed, there are optimum tariffs, which maximize a country's gain from trade, and whose level depends on the price elasticities of the foreign supply of imports and the foreigners' demand for exports, just as the monopolist's profit maximizing profit margin depends on the price elasticity of demand he faces (Scitovsky 1942).

Tariffs, like monopoly pricing, redistribute income in favour of those imposing them in a way that inflicts a greater loss on those hurt than the gain they secure for those favoured. For that reason, it is important to prevent competitive tariff impositions and increases, whereby each country retaliates in self-defence to the tariffs imposed by others, and so contributes to a general impoverishment of all or almost all, due to the all-round reduction of international specialization and of the gain it generates. Yet, that happened during the 1930s depression; and it can easily happen when each country believes itself to have a small enough share in world trade to erect or raise tariffs unpunished and retaliation is effective or believed to be effective in recapturing some of the lost gain of the retaliating country. Free trade therefore is not a stable situation, unless imposed by a dominant country, such as Great Britain in the nineteenth century or the United States during the period following World War II (Scitovsky 1942; Kahn 1947). What happens when free trade is not enforced, which countries gain, which lose from tariffs and retaliatory tariffs, and what is the nature of the path and final outcome of competitive trade restrictions has received considerable attention in the theoretical literature (Kaldor 1940; Scitovsky 1942; Graaff 1949–1950; Johnson 1953, 1954; Gorman 1957, 1958), but is too complex to summarize here.

Also, the subject has remained a theoretical exercise and faded into the background. Yet, it has two aspects that, though largely overlooked in the literature, deserve mention here. One is that tariffs can exploit a country's monopoly or monopsony position in world markets *only* if that is not already exploited by private firms within the country. The other is that an import duty can be used as countervailing power to prevent a country's being exploited by a foreign exporter's use of *his* monopoly power.

A country's only large producer of an exportable product or its single importer of a foreign product enjoys, of course, the same monopoly or monopsony position in world markets as does the country as a whole. Accordingly, he can, and usually does, exploit that position to his own – as well as to his country's – advantage by setting the profit-maximizing monopoly or monopsony price. The same is approximately true also if, instead of a single monopolist, a few large firms act in open or tacit oligopolistic collusion in setting monopoly prices. When they do that, tariffs for the purpose of exploiting the country's bargaining position in world markets are not only redundant but harmful, because, added to a producer's monopoly (or subtracted from an importer's monopsony) price, they are liable to push the foreign price beyond its profit-maximizing level, thereby inflicting a loss on domestic exporters or importers that exceeds the government's tariff revenue. In short, tariffs and monopolistic profit margins can substitute one for another, complement each other, but cannot be used to exploit the same monopoly or monopsony position twice over.

That explains, for example, why export tariffs and other export restrictions have been imposed almost exclusively on primary products and only in countries where those are grown by many small growers under competitive conditions. Export duties on coffee and the Ghanaian State monopoly for the export of cocoa are the obvious examples. The industrial countries, which export manufactures, have no need for export duties to exploit their monopoly position in world markets, because the large manufacturers of their exportables are usually able to charge monopoly prices on their own, thus making export duties redundant.

The same argument also explains why Britain practised and preached free trade up to the end of the nineteenth century. Her heavy manufactures were produced and exported by large, monopolistic firms, her light manufactures (textiles), though

produced competitively, were exported by large wholesale merchants, and some of her primary-product imports were also handled by large British firms, most of them able to set prices that exploited their foreign and domestic monopoly positions alike and rendered tariffs superfluous.

We come now to the use of an import duty to offset a foreign exporter's monopoly and diminish or eliminate his monopoly profits. Ross Shepherd has shown (Shepherd 1978) that a variable import duty which varies, and is *expected to vary,* directly with the foreign price of an imported good, raises the country's apparent price elasticity of demand for that import and correspondingly reduces its manufacturer's monopoly power and with it his profit maximizing price. Indeed, under constant cost conditions, a suitable duty will leave unchanged both the volume imported and the domestic price paid for it by domestic consumers, while expropriating the foreign exporter's monopoly profit. Abba Lerner, who seems to have arrived at the same conclusion independently, advocated imposing such a variable duty (which he called 'extortion tax') on oil imports, thereby creating an incentive for OPEC's members to break ranks by reducing price (Lerner 1980).

In closing, it is worth noting some similarities and differences between the imposition of tariffs and devaluation. A uniform ad valorem duty on all imports combined with a uniform ad valorem subsidy (negative duty) of the same magnitude on all exports is identical to a devaluation of that magnitude in its effects on the balance of trade but leaves unchanged all other international transactions and financial obligations. For that reason, countries anxious not to increase the burden on domestic debtors of foreign debt denominated in foreign currencies have used such and similar policies as means of improving their balance of trade in preference to devaluation. Also, since devaluation worsens a country's terms of trade when the foreign demand for some of its exports is very inelastic, it may be combined with a duty or other restraint on those of its exports (usually primary products), thereby to prevent the deterioration of its terms of trade, or import restriction may be substituted for devaluation.

See Also

▶ Optimal Tariffs

Bibliography

Bickerdike, C.F. 1906. The theory of incipient taxes. *Economic Journal* 16: 529–535.
de V. Graaff, J. 1949. On optimum tariff structures. *Review of Economic Studies* 17: 47–59.
Gorman, W.M. 1958. Tariffs, retaliation, and the elasticity of demand for imports. *Review of Economic Studies* 25: 133–162.
Johnson, H.G. 1954. Optimum tariffs and retaliation. *Review of Economic Studies* 21: 142–153.
Kahn, R.F. 1947. Tariffs and the terms of trade. *Review of Economic Studies* 15: 14–19.
Kaldor, N. 1940. A note on tariffs and the terms of trade. *Economica* 7: 377–380.
Lerner, A.P. 1980. OPEC – A plan – If you can't beat them, join them. *Atlantic Economic Journal*, Sept, 1–3.
List, F. 1841. *The national system of political economy.* Trans. S.S. Lloyd. London: Longmans, Green & Co., 1885.
Little, I.M.D., T. Scitovsky, and M.K. Scott. 1970. *Industry and trade in some developing countries: A comparative study.* London: Oxford University Press.
Scitovsky, T. 1942. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9: 89–110.
Shepherd, A.R. 1978. *International economics: A micro-macro approach*. Columbus: Charles E. Merrill.

Tarshis, Lorie (1911–1993)

D. E. Moggridge

Keywords
Aggregate supply function; Keynesianism; Tarshis, L.; Underinvestment

JEL Classifications
B31

Tarshis was born in Toronto, Canada, on 22 March 1911. After a commerce degree at the University of Toronto, he went to Trinity College, Cambridge, where he took a BA in 1934 and a Ph.D.

in 1939. His years in Cambridge, 1932–6, which coincided with the emergence of Keynes's *General Theory,* shaped much of his subsequent professional life. His notes for Keynes's annual series of eight lectures on his work in progress for the years 1932–5 have become an important source for those interested in tracing the evolution of Keynes's views. The two Cambridge revolutions of the 1930s, Keynes's and imperfect competition, focused the analysis of his Ph.D. dissertation, 'The Distribution of Labour Income'. From this came two classic articles in 1938 and 1939 which, along with a contemporaneous piece by John Dunlop (1938), forced Keynes to reconsider his generalization that real and money wages moved inversely over the trade cycle and its implications for the assumption of perfect competition that underlay the analysis of the book (Keynes 1939).

By then Tarshis had moved to the United States, first to Tufts University (1936–9, 1942–6) and subsequently to Stanford (1946–71). While at Tufts, along with his Cambridge classmate R.B. Bryce, he played a significant role in spreading Keynes's ideas among the Harvard community of economists. Then in 1938 he participated with several other economists in the manifesto *An Economic Program for American Democracy.* Only seven of them eventually signed it – R.V. Gilbert, G.H. Hildebrand Jr., A.W. Stuart, M.Y. and P.-M. Sweezy, Tarshis and J.D. Wilson – the government or other connections of the rest preventing them from doing so. The *Program* was 'Keynesian in analysis, stagnationist in diagnosis and all-out in prescription', and was 'instrumental' in driving home to New Deal Washington the need for more spending to overcome the fatal flaw of contemporary capitalism, underinvestment (Stein 1969, pp. 165–7). His move to Stanford coincided with another effort at Keynesian persuasion, *The Elements of Economics,* the first unashamedly Keynesian introductory textbook. Dogged by controversy over its supposed 'left wing' views, it was much less successful than the slightly later competing text of Paul Samuelson.

During the subsequent 40 years, despite his heavy teaching commitments where he probably left his greatest mark, Tarshis continued to publish regularly. His contributions related to international finance, the microeconomics of Keynes (most notably the aggregate supply function) and contemporary policy issues.

Selected Works

1938a. Real wages in the United States and Great Britain. *Canadian Journal of Economics and Political Science* 4: 362–376.
1938b. (With R.V. Gilbert et al.) *An economic program for American democracy.* New York: Vanguard Press.
1939. Changes in real and money wages. *Economic Journal* 49: 150–154.
1947. *The elements of economics.* Boston: Houghton Mifflin.
1979. The aggregate supply function in Keynes's *General Theory.* In *Economics and human welfare: Essays in honour of Tibor Scitovsky,* ed. M.J. Boskin. New York: Academic.
1984. *World economy in crisis.* Toronto: Lorimer.

Bibliography

Dunlop, J.T. 1938. The movement of real and money wage rates. *Economic Journal* 48: 413–434.
Keynes, J.M. 1939. Relative movements of real wages and output. *Economic Journal* 49: 34–51.
Stein, H. 1969. *The fiscal revolution in America*. Chicago: University of Chicago Press.

Taste-Based Discrimination

Kerwin Kofi Charles and Jonathan Guryan

Abstract

Economists typically account for differences in economic outcomes between ethnic groups with explanations having to do with differences in skill, explanations emphasizing informational problems associated with accurately assessing skill, as in statistical discrimination models, or explanations that rely on the

presence of prejudice, the key element of taste-based discrimination models. This article defines taste-based discrimination and briefly outlines the economics of associated models. It discusses empirical implications of these models, and reviews empirical tests from the literature. It speculates about possible avenues for future research likely to enrich the insights forthcoming from the standard taste-based model. Although this article focuses on discrimination arising from racial prejudice, taste-based discrimination subsumes negative preferences towards groups of individuals of many alternative types, including different age, gender or religious groups.

Keywords

Discrimination; Prejudice; Race; Segregation; Taste-based model; Wage levels

JEL Classifications

J01; J7; J71; J31; J15

Becker's Taste-Based Discrimination Models

Individual Preferences

In his seminal work, Becker (1971; 1st edn 1957) formally demonstrated how negative racial feeling, or prejudice, on the part of individual members of a majority group (here described as whites) could be related in a market environment to negative outcomes for members of a discriminated-against group (here described as blacks). He chose a representation of racial prejudice which, in addition to being precise and intuitively appealing, lent itself readily to tractable representation in an economic model. Prejudice in Becker's framework is represented as an aversion to cross-racial (or more generally cross-group) contact. Since this aversion renders cross-racial interactions psychically costly, the strength of an agent's aversion can also be thought of as the price the person would be willing to pay to avoid the interaction. Becker studies the effect of prejudice among three distinct types of white agents – employers, employees and customers.

Given the representation of racial prejudice, it follows straightforwardly that, in their market interactions that involve blacks, prejudiced agents act *as if* the relevant price mediating that interaction is the actual market price plus an amount determined by the agent's level of prejudice. Thus prejudiced employers with a taste for discrimination, d_i, deciding about hiring black workers view themselves as paying not the market wage w they pay to white workers but rather the price $w + d_i$. (This functional form assumes the disutility of interaction is linear in the number of black employees. Other functional forms are of course possible.) When the prejudiced person is an employee, holding a taste for discrimination Δ_i and contemplating a wage offer of $\hat{w}$ to work at a firm alongside black workers, he views himself to be working for a wage of $\hat{w} - \Delta_i$, rather than the wage $\hat{w}$ he would consider himself to be receiving were all his co-workers white. Finally, in the third example studied by Becker, a prejudiced customer with a taste for discrimination κ_i views himself as paying $p + \kappa_i$ per unit for goods sold by black sellers, rather than the market price p he would consider himself to be paying were the sellers white. It should be clear that the parameters d_i, Δ_i and κ_i each reflect the disutility that a prejudiced agent receives from interacting with blacks.

Market Implications

What does individual prejudice as represented above imply about equilibrium wages and prices? Keeping closely to Becker's original presentation, we briefly describe how tastes interact in a market setting, where there is optimizing behaviour and competition, to determine the level of market discrimination for each of the three models.

Employer Discrimination

If black (b) and white (a) workers are equally productive and perfect substitutes in production, an employer i will have utility given by

$$U_i = f(K, L_a + L_b) - w_a L_a + w_b L_b - d_i L_b$$

where $f(\cdot)$ is a constant returns to scale production function, w_j is the market wage paid to workers from group $j \in (a, b)$, K is capital, and L_j is the number of workers hired from group j. Taste for discrimination, $d_i \geq 0$, among all employers varies according to some arbitrary distribution Ω.

Since in this model black and white workers are perfect substitutes in production, each employer simply hires the type of worker who is less costly, at the margin, *to him*. Thus an employer hires only black workers if $w_b + d_i < w_a$, and he hires only white workers if the strict inequality is reversed. Notice that an employer's workforce is strictly segregated by race, unless his racial prejudice d_i is such that $w_b + d_i = w_a$.

What is the equilibrium in the short run, when the number and size of firms are fixed? Imagine a central planner choosing black and white wages and allocating black and white workers to employers so that the markets for both types of workers clear. The planner allocates black workers to the least prejudiced employers: that is, to those with $d_i = 0$ first and then, if necessary, to those with the lowest values of d_i. If the distribution of d_i is smooth, the last employer to be allocated a black worker must be indifferent between hiring black and white workers. In equilibrium, less prejudiced employers hire blacks, more prejudiced employers hire whites, and the equilibrium black–white wage gap $w_a^* - w_b^*$ is equal to the prejudice of the employer who is indifferent between hiring blacks and whites at the equilibrium wage, or d_i^*.

The model sharply distinguishes individual prejudice from market discrimination. In particular, the equilibrium black–white wage gap is not determined by the average level of prejudice among employers, but by the prejudice of a marginal discriminator. Even in a market in which some employers are prejudiced, there need be no racial wage gap in equilibrium provided there are non-prejudiced employers to hire all the black workers. Notice also that since prejudiced employers and black workers have an incentive to avoid interacting, there is market pressure towards segregation, so that segregation and observed discrimination are effectively substitutes. The more segregation there is in equilibrium, the smaller is the wage gap.

In general, the equilibrium black–white wage gap increases as the prejudice of the marginal discriminator increases – either because the fraction of the workforce that is black increases or because of changes in the distribution of prejudice among all employers. In the first case, the presence of more blacks in the market means that market clearance requires that blacks be allocated to ever more prejudiced employers at the margin. In the second case, higher levels of prejudice in *the part of the distribution from which the marginal employer is likely to be drawn* will increase market discrimination. Since blacks represent a small minority in most markets, the sorting that characterizes the equilibrium guarantees that only higher prejudice among those employers in the left tail of the distribution (below the median) should lower black wages; higher prejudice among the most prejudiced employers in the market should have no effect on the equilibrium wage gap since the marginal employer is very unlikely to be drawn from among these persons.

Employee Discrimination

As discussed above, a prejudiced employee behaves as if the wage he is offered by a firm with black workers were the actual offered wage $\hat{w}$ minus the disutility he gets from interacting with blacks at work, Δ_i. As in the employer discrimination model, market forces generate a tendency toward segregation. In the employee discrimination case, because of the preferences of their workers, employers have an incentive to segregate their workforces. An employer who hires both black and prejudiced white workers is forced to pay a premium to the whites to induce them to work for him. He does not pay that premium if his workforce is all the same race. Each firm therefore prefers to employ either only white workers or some combination of black and unprejudiced white workers.

If, in equilibrium, firms are able to segregate perfectly by race, there will be no equilibrium wage gap. Only if there are impediments to perfect segregation, large enough to ensure that blacks

work with prejudiced co-workers, can employee prejudice lead to wage discrimination. In the likely event these frictions are such that more prejudiced employees are especially unlikely to work with black co-workers, reductions in segregation lead to increases in the racial wage gap.

Customer Discrimination

In the third type of taste-based model discussed by Becker, some customers care about the race of workers producing the goods they purchase. They consequently regard the price they pay for goods made by a firm with l_b black workers to be the charged price, P_b, plus their disutility which increases with the number of black inputs into production, or $\kappa_i l_b$. Since all employers are unprejudiced, and since the profits of firms hiring black and white workers must be the same in equilibrium, any per unit price difference in equilibrium must be reflected in a difference in wages paid to black and white workers. (P_a is the price of a good produced exclusively by white workers.) If there is a price difference in the good, the most prejudiced customers will buy goods produced by whites, and the least prejudiced customers will buy goods produced by blacks. The marginal discriminator is that consumer who is indifferent between buying goods made by blacks and those made by whites, given his level of prejudice and the equilibrium prices charged. If there are enough unprejudiced customers relative to the number of blacks in the market, there will be no difference in prices between the two types of goods, and no difference in equilibrium wages by race. If this condition does not hold, then there will be a racial wage gap.

Long-Run Implications

Traditional View

Much early discussion and criticism about taste-based models centred on the nature of the long run, when firms can freely enter and exit from the market. The employer version of the taste-discrimination model has historically been the focus of this criticism, as it is in this particular prejudice model that the long-run implications appear, at first blush, to be most troubling for the standard model.

To see the essence of the criticism of the employer prejudice model, note that if there is an equilibrium wage gap in the short run, employers that hire only white workers have higher labour costs than do firms that hire only black workers. Since workers are equally productive by assumption, non-discriminating firms earn greater profits. The return to capital is thus different across firms, and in the long run, capital should flow to those firms with the highest return – those with the least prejudiced employers. With a constant returns to scale production function, this mechanism continues until all employers with $d_i > 0$ shut down and leave the market. In the long-run equilibrium, no prejudiced employer survives and any wage gap caused by prejudice is eliminated. Becker himself originally outlined this argument in his original work, and it was famously and forcefully repeated later by Arrow (1972). The influence of Arrow's criticism was such that it probably discouraged work on taste-based prejudice models, relatively few of which have appeared subsequently in the theoretical literature. At the same time it probably encouraged work on statistical discrimination, which is today the dominant paradigm for the theoretical study of discrimination. Indeed, the earliest versions of statistical discrimination model were presented by Arrow (1972) in the same paper in which his criticism was lodged, followed quickly by Phelps's (1972) seminal work.

Recent Theoretical Work

In recent years a number of authors have presented employer prejudice models of discrimination that modify Becker's original assumptions about competition or the wage-setting process, or introduce job search frictions to resurrect the prediction of long-run racial wage gaps resulting directly from taste-based discrimination.

In an important paper, Black (1995) relaxes the perfect competition assumption, introducing costly job search into a model in which some employers refuse to hire black workers and others are non-discriminatory. Without perfect information, black workers know that some fraction

of the employers they randomly encounter while searching would refuse to hire them. Non-discriminatory employers consequently enjoy local monopsony power over their black workers, in the sense that reductions in their wage will not lead all such workers to leave. In equilibrium, the presence of employers that refuse to hire blacks causes blacks to receive, from those employers that *do* hire them, lower wages than otherwise identical whites.

More recently Lang et al. (2005) described a model in which employers post binding wage offers. Because firms cannot base wage offers on the race of the worker, discriminatory employers exercise their preferences by choosing to hire based on race. Since job search is costly, black workers choose to apply to firms at which they believe they have a good chance of being hired. These firms are those that post low wage offers. In equilibrium, there is both segregation and a wage gap.

Our own recent work (Charles and Guryan 2007, 2008) calls into question the basic logic of Arrow's critique of the Becker model. The conclusion that discriminatory employers are driven to shut down as a result of competition with non-discriminatory employers is based on an assumption about employers' alternative labor market opportunities, which may be as a worker at a firm. By its reliance on zero profit as the condition for shutting down, this conclusion is implicitly based on the assumption that an employer's outside option is valued only according to the monetary wage paid at that firm, and not according to the race of his potential co-workers. If instead it is assumed that a prejudiced employer takes his distaste for racial contact along with him to his role as a co-worker, then the conclusion is different. Whether a prejudiced employer shuts down then depends on his likelihood of finding a job that does not involve contact with black co-workers. It follows that the ability of the market to segregate is the key mediator between individual discriminatory tastes and market discrimination. In a model with racial preferences that are portable across roles (employer or worker) and in which there is an impediment to segregation, racial wage gaps can persist in the long run even in the face of perfect competition and free entry and exit.

Interestingly, we might think of the model described in Charles and Guryan (2007, 2008) as one in which the choice of whether to be an employer or an employee is endogenized. In this case, Becker's worker discrimination model is essentially a generalization of his employer discrimination model.

The predictions of this model are essentially the same as those from Becker's short-run employer discrimination equilibrium.

Empirical Assessment of Taste-Based Prejudice Models

There is a vast empirical literature devoted to studying racial gaps in economic outcomes, but very little of that work can be said to test directly the implications of taste-based models. Part of the reason is that it is difficult to establish that observed racial wage differences are truly the result of any form of discrimination and not of unmeasured skill or ability differences across the groups. The suspicion that unmeasured skill differences account in part for observed wage and other differences by race is strengthened by results, such as those from Neal and Johnson (1996), showing that adding rarely used controls like test scores to wage regressions results in a significant reduction in the amount of the wage gap that is unexplained. Even when it can be reasonably argued that the unexplained gaps are unlikely to be the result of some measure of skill for which the researcher has failed to control, there remains the problem that such results are consistent with forms of discrimination that have nothing at all to do with taste-based prejudice. This is true even of the important and carefully done audit studies in various markets, in which blacks and whites, or their resumes, are sent at random to different employers or firms in analysts, attempts to identify differential treatment. (See for example Heckman 1998; Bertrand and Mullainathan 2004.)

Charles and Guryan (2008) test the predictions of Becker's model empirically, collecting data on

explicit measures of prejudice from the General Social Survey (GSS) to construct indices of the distribution of individual discriminatory tastes. They then compare various measures of prejudice at the labour market (in this US example, state) level, and compare these with the measured black–white wage gap in the market. The results are remarkably supportive of Becker's model. Black–white wage gaps are larger in states with a higher fraction of a black workforce. Racial wage gaps are larger in more prejudiced states, but the relationship holds in a particular way. Black relative wages are negatively related to the degree of prejudice in the left tail of the prejudice distribution, but not to variation in the median or right tail. This finding is consistent with the sorting mechanism that Becker describes, and that tends to make the marginal discriminator someone less prejudiced than the average person. Further supporting the view that the market's ability to segregate is a key mediator between individual tastes and market outcomes, the study found that black–white wage gaps are larger in states where there is more integration of – more contact between – blacks and whites in the workplace.

Areas for Future Work and Conclusion

Despite the historical importance of taste-based models in the economics literature on discrimination, we believe there are many areas for future theoretical and empirical work. We briefly discuss only a few of these, listing them in the form of questions.

What Precisely Is Taste-Based Prejudice?

Becker represents racial prejudice as distaste for interaction, and virtually all subsequent prejudice models have followed Becker's lead. However, racial prejudice might manifest itself in various other ways, with possible important implications for market outcomes. Prejudice might, for example, affect information processing. In particular, racial prejudice might cause people infected by it not to update negative stereotypes about members of a group, even in the face of contradictory evidence. This representation of prejudice is closely related to work by authors like Loury (2002) and Coate and Loury (1993) about the formation of stereotypes, and to Myrdal's (1944) work on 'vicious cycles'. Much work remains to be done with this representation of prejudice. Nor has much work been done with prejudice represented as a preference for one's own group (nepotism) rather than an aversion to interactions with another group. (See Goldberg 1982 for an important exception.)

A massive literature in social psychology examines the formation of prejudice and stereotypes, including the question of whether prejudice is a preference towards members of an in-group or dislike of members of an out-group. (It would be impossible to summarize that entire literature in this article, but for a good overview, the interested reader should see Fiske 1998. An important early study is Allport 1954.)

The interesting experimental work of Tajfel et al. (1971) suggests that agents care about the utility of their 'own' group, but may care more about actions that maximize the *difference* between their group's outcomes and those of the out-group. Exploring the implications of these and other insights in formal economic models would vastly enrich our understanding of the effect of prejudice.

From Where Does Racial Prejudice Come?

Are group-based preferences instinctual or learned? What are the root causes of prejudice? Is prejudice the result of ignorance or unfamiliarity? The answers to questions like these are important for designing policies aimed at reducing discrimination, or even the prevalence of the tastes themselves. Research in psychology and sociology, such as that by Tajfel et al. (1971) and the famous 'Robbers' Cave' study by Sherif et al. (1961/1988), suggests that the entire notion of in- and out-groups, with associated negative and positive feelings, might even arise among people randomly assigned to groups. At the same time, some of this research suggests that in- and out-group sentiments are especially strong when there is competition over scarce resources. Future work by economists might incorporate these insights to how prejudice could arise

endogenously from, say, residential segregation. Or, future work might assess whether, following sectoral reallocation or local demand shocks, there is a change in the extent to which prejudice varies with competition over scarce jobs.

Is Prejudice Conscious?

Most analyses of prejudice in economics assume that the prejudiced agent is aware of his prejudicial sentiments. But research suggests that one way that people might conserve their limited cognitive resources is to group objects into categories and then to generate summary beliefs for those groups rather than a separate one for each individual (see e.g. Allport 1954; Tajfel 1981; Taylor 1981). Humans may therefore develop prejudicial beliefs of which they are not overtly conscious. An example of a test of such subconscious prejudice is the Implicit Attitudes Test (Banaji and Greenwald 1995; Greenwald et al. 1998). Results in the lab (Wittenbrink et al. 1997) and in the field (Price and Wolfers 2007) support the possibility that automatic psychological responses contribute to discriminatory actions. Further investigation of whether prejudice is subconscious and the associated empirical consequences of that being the case would be an obvious useful area for future work.

In addition to answering questions like these, there is clearly a need to subject the various prejudice models to additional empirical tests. Charles and Guryan (2008) is the only paper of which we are aware that tests the basic predictions of the employer discrimination model about the relationship between equilibrium wage gaps and the distribution of prejudice. Clearly there should be more of this type of work, across different areas and over different time periods. To that end, there might be great benefit to economists from collecting data on prejudice themselves rather than relying on data collected by others. For example, no prejudice measure of which we are aware elicits from respondents an answer to the question how much they would be willing to pay to avoid interacting with members of a particular group. Economists collecting such data would naturally cast their prejudice questions in this form, allowing for an almost exact translation between the theoretical construct of interest to economists and the measure used to study it in the data. Similarly, given the central theoretical importance of the nature and frequency of cross-group interaction within firms to observed racial wage gaps in taste-based models, there is likely to be a great deal to be learned from empirical analyses that jointly study wage discrimination and segregation. Here too there would seem to be substantial returns from the collection of new data, as the available data on actual interactions within firms is coarse and available in very few data sources.

Taste-based models were the first models of discrimination written down by economists. While these models almost certainly cannot account for all of the large and durable differences observed in the market across racial, gender and other dimensions, they nonetheless yield valuable insights about why putatively unproductive traits like race or gender might be correlated systematically with worse market outcomes. Although most work on discrimination has focused on explanations like statistical discrimination or systematic differences in unmeasured skill, there is evidence of a resurgence in interest in taste-based models, given the appearance of a number of papers in the recent literature. In our view, renewed interest in, and investigation of, taste-based models will inevitably lead to a richer understanding of the nature and reasons for differences in economic outcomes in the population.

See Also

- ▶ Anti-Discrimination Law
- ▶ Labour Market Discrimination

T

Bibliography

Allport, G.W. 1954. *The nature of prejudice*. Reading: Addison-Wesley.

Arrow, K. 1972. Some mathematical models of race in the labor market. In *Racial discrimination in economic life*, ed. A.H. Pascal, 187–204. Lexington: Lexington Books.

Banaji, M.R., and A.G. Greenwald. 1995. Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology* 68: 181–198.

Becker, G.S. 1971. *The economics of discrimination*, 2nd ed. Chicago: University of Chicago Press.
Bertrand, M., and S. Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94(4): 991–1013.
Black, D.A. 1995. Discrimination in an equilibrium search model. *Journal of Labor Economics* 13(2): 309–334.
Charles, K.K., and J. Guryan. 2007. *Prejudice and the economics of discrimination*, NBER Working Paper No. 13661. Cambridge, MA: NBER.
Charles, K.K., and J. Guryan. 2008. Prejudice and wages: An empirical assessment of Becker's the economics of discrimination. *Journal of Political Economy* 116(5): 773–809.
Coate, S., and G.C. Loury. 1993. Will affirmative-action policies eliminate negative stereotypes? *American Economic Review* 83(5): 1220–1240.
Fiske, S.T. 1998. Stereotyping, prejudice and discrimination. In *The handbook of social psychology*, ed. D.T. Gilbert, S.T. Fiske, and G. Lindzey. New York: Oxford University Press.
Goldberg, M.S. 1982. Discrimination, nepotism, and the long-run wage differential. *Quarterly Journal of Economics* 97(2): 307–319.
Greenwald, A.G., D.E. McGhee, and J.K.L. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74: 1464–1480.
Heckman, J.J. 1998. Detecting discrimination. *Journal of Economic Perspectives* 12(2): 101–116.
Lang, K., M. Manove, and W. Dickens. 2005. Racial discrimination in labor markets with posted wage offers. *American Economic Review* 95(4): 1327–1340.
Loury, G.C. 2002. *The anatomy of racial inequality.* Cambridge, MA: Harvard University Press.
Myrdal, G. 1944. *An American dilemma: The Negro problem and American democracy.* New York: Harper & Brothers.
Neal, D.A., and W.R. Johnson. 1996. The role of premarket factors in black–white wage differences. *Journal of Political Economy* 104(5): 869–895.
Phelps, E.S. 1972. The statistical theory of racism and sexism. *American Economic Review* 62: 659–661.
Price, J., and J. Wolfers. 2007. *Racial discrimination among NBA referees*, NBER Working Paper No. 13206. Cambridge, MA: NBER.
Sherif, M., O.J. Harvey, B.J. White, W.R. Hood, and C.W. Sherif. 1961/1988. *The Robbers Cave experiment: Intergroup conflict and cooperation*. Middletown: Wesleyan University Press
Tajfel, H. 1981. *Human groups and social categories: Studies in social psychology.* Cambridge, MA: Cambridge University Press.
Tajfel, H., C. Flament, M. Billig, and R. Bundy. 1971. Social categorization and intergroup behavior. *European Journal of Social Psychology* 1(2): 149–178.
Taylor, S.E. 1981. A categorization approach to stereotyping. In *Cognitive processes in stereotyping and intergroup behavior*, ed. D.L. Hamilton, 83–114. Hillsdale: Elrbaum.
Wittenbrink, W., C.M. Judd, and B. Park. 1997. Evidence for racial prejudice and the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology* 72: 262–274.

Tâtonnement and Recontracting

Takashi Negishi

Keywords

Auctioneer; Barter; Edgeworth, F. Y.; Exchange; General equilibrium; Gross substitutability; Hicks, J. R.; Kinked demand curve; Law of indifference; Limit theorem; Marshall, A.; Money; Neutrality of money; Non-recontracting models; Non-tâtonnement models; Numéraire; Recontracting; Sticky prices; Tâtonnement; Tâtonnement and recontracting; Walras, L.; Walras's Law

JEL Classifications

D5

In the current theory of general economic equilibrium, recontracting and tâtonnement (a French word meaning 'groping') are used interchangeably to denote a simplifying assumption that no actual transactions, and therefore no production and consumption activities, take place at disequilibria when prices are changed according to the law of supply and demand (Kaldor 1934; Arrow and Hahn 1971, pp. 264, 282). Historically speaking, however, this usage is somewhat confusing, since recontracting is originally due to Edgeworth, who developed it in a direction different from that in which Walras developed his tâtonnement (Walker 1973).

Though different interpretations are given as to whether Walras explicitly excluded disequilibrium transactions from the beginning (Patinkin 1956, p. 533; Newman 1965, p. 102; Jaffé 1967, 1981), it is clear that Walras developed his theory of

tâtonnement so as to exclude such transactions. To do this there are at least three methods of tâtonnement. First, we may assume that price-taking traders facing market prices cried by the *auctioneer* reveal their plans of demand and supply to the auctioneer but do not make any trade contract among themselves until the auctioneer declares that equilibrium is established. Alternatively, traders may be assumed to make trade contracts (Walras 1926, p. 242, suggested the use of tickets when production is involved) but recontract is assumed always to be possible, in the sense that contract can be cancelled without consent of the other party if market prices are changed. Finally, the effect of past contracts can be nullified by offering new demands (supplies) to offset past supplies (demands), even if it is assumed that past contracts are effective and would be carried out at the current prices when the equilibrium is established (Morishima 1977, pp. 28–30). Since any changes in prices make the contract unfavourable to one of the parties which then wishes to cancel the trade contract, there is no difference between the three methods of tâtonnement in the behaviour of demand, supply and prices. Recontracting in this sense of tâtonnement is, however, quite different from that developed in Edgeworth's theory of recontract.

We shall start by the consideration of why this assumption of tâtonnement is necessary for the Walrasian theory of general equilibrium, which is the foundation of neoclassical economic theory. The reason lies in the structure of Walrasian economics, dichotomized between real and monetary theories. Then we analyse formal models of tâtonnement including the original one due to Walras and the modified version developed in modern theories of general equilibrium. It is followed by our assessment of the theoretical achievements and empirical relevance of Walrasian tâtonnement economics. Edgeworth's theory of recontract is reviewed in its relation to the Walrasian theory of tâtonnement. Finally, an evaluation is made on the recent studies of tâtonnement and recontracting, to show in which direction further progress should be made.

Walras (1874–1877) insisted that complicated phenomena can be studied only if the rule of proceedings from the simple to the complex is always observed. To understand the fundamental nature of Walrasian economics, it is convenient to make (as did Hicks 1934) a comparison of Walrasian and Marshallian ways of applying this rule to the study of complicated economic phenomena. Both Walras and Marshall (1890) start with a very simple model of an economy and then proceed to more complex models. There is an important difference, however, between Walrasian general equilibrium analysis and Marshallian partial equilibrium analysis.

Walras first decomposes a complicated economy of the real world into several fundamental components like consumer-traders, entrepreneurs, consumers' goods; factors of production, newly produced capital goods, and money. He then composes a simple model of a pure exchange economy by picking up a very limited number of such components, that is, individual consumer-traders and consumer's goods, disregarding the existence of all other components. Travel from this simple model to the complex proceeds by adding one by one those components so far excluded, that is, entrepreneurs and factors of production first, then newly produced capital goods, and finally money. In this journey each intermediate model, enlarged from a simpler one and to be enlarged into a more complex one, is still a closed and self-compact logical system. However, each of them is as unrealistic as the starting model, with the exception of the last, into which all the components of a real world economy have been introduced.

Marshall on the other hand studies a whole complex of a real world economy as such. Of course, he also simplifies his study at first by confining his interest to a certain limited number of aspects of the economy. But he does it not by disregarding the existence of other aspects but by assuming that other things remain equal. In this sense most of Marshall's models of an economy, though realistic, are open and not self-sufficient, since some endogenous variables (that is, the 'other things') remain unexplained and have to be given exogenously.

The simplest model of Walrasian economics is that studied in the theory of exchange, where goods to be exchanged among individual

consumer-traders are assumed simply to be endowed to them and not considered as produced at cost. There exist no production activities in this hypothetical world. The corresponding simplest model of Marshall is that of the market day, in which goods to be sold are produced goods, although the amount available for sale is, for the time being, assumed to be constant. Production does exist in this temporary model, though the level of output is assumed to be unchanged. In that Walrasian model which includes production capital goods are introduced as a kind of factor of production but investment (that is, the production of new capital goods) simply does not exist. On the other hand, in Marshallian short-run theory, which is also a theory of production, investment is actually undertaken though the amount of currently available capital is given. In all of the Walrasian models of exchange, of production and of credit and capital formation there exists no money at all, until it is finally introduced in the theory of circulation and money. In Marshallian models on the other hand money exists from the beginning, though its purchasing power is sometimes assumed to be constant.

In other words, Marshallian theories correspond respectively to special states of the real world economy. The market day (temporary) and short-run models are just as realistic as the long-run model, where capitals are fully adjusted. Thus Marshallian models are practically useful to apply to what Hicks (1934) called particular problems of history or experience. On the other hand, Walrasian models are in general not useful for such practical purposes. They are designed to show the fundamental significance of such components of the real world economy as entrepreneurs and production, investment and the rate of interest, inventories and money, and so on, by successively introducing them into simple models which are then developed into more complex ones. Walras' theoretical interest was not in the solution of particular problems but in what Hicks called the pursuit of the general principles which underlie the working of a market economy.

From our standpoint we must emphasize that all exchanges have to be nonmonetary (that is, direct exchanges of goods for goods) in all the Walrasian theories of exchange, production and capital formation and credit, since money has not yet been introduced. Relative prices (including the rate of interest) and hence consumption and production activities are determined in non-monetary real models without using money, while the role of the model of circulation and money lies only in the determination of the level of absolute prices by the use of the money (Morishima 1977, ch. 11; Negishi 1979, ch. 2). Thus Walrasian economics is completely dichotomized between non-monetary real theories and monetary theory, in the sense that all non-monetary real variables are determined in the former and money is neutral, that is it does not matter for the determination of such variables. 'That being the case, the equation of monetary circulation, when money is not a commodity, comes very close, in reality, to falling outside the system of equations of (general) economic equilibrium' (Walras 1926, pp. 326–7).

In each of his non-monetary theories Walras tried to show the existence of a general equilibrium in its corresponding self-compact closed model. General equilibrium is of course a state in which not only each individual consumer-trader (entrepreneur) achieves the maximum obtainable satisfaction (profit) under given conditions but also demand and supply are equalized in all markets. In a large economy, how can we make such a situation possible without introducing money? What kind of process of exchange should we consider in order to establish a general equilibrium without using money? Even in the most simple case of an exchange economy, it seems in general almost impossible to satisfy all individual traders by barter exchanges, unless mutual coincidence of wants accidentally prevails everywhere. Walras ingeniously solved this difficulty by his famous tâtonnement, a preliminary process of price (and quantity) adjustment which precedes exchange transactions and/or effective contracts.

Suppose that all the individual consumer-traders and entrepreneurs meet in a big hall. Since all of them are assumed to be competitive price takers it is convenient to assume (though Walras himself did not do so explicitly) the existence of an *auctioneer* whose only role is to determine prices. At the start the auctioneer calls all

prices (including the price of a bond) at random. Individual consumers and entrepreneurs make decisions on the supply and demand of all goods, factors of production and of the bond, assuming that the prices cried by the auctioneer are fixed and that whatever amount they wish can actually be supplied and demanded at these prices. If total demand equals total supply for every good (including the factors of production and the bond) exchange takes place (or contracts are made) at these prices, and the problem is solved.

Generally, however, this will not be the case, in which event no exchange transaction should take place at all, even for a good for which total demand is equal to total supply, and every mutually agreed contract should be cancelled. The auctioneer cancels the earlier prices, which failed to establish a general equilibrium, and calls new prices by following the law of supply and demand, that is, raising (lowering) the price of each good for which the demand is larger (smaller) than the supply. The same procedure is repeated until general equilibrium is established. Actual exchange transactions take place and enforceable contracts are made only when every party can actually realize its plan of demand and supply.

Prices change in the process of tâtonnement and it is generally impossible for a single trader to purchase or sell whatever amount he wishes at going prices., each trader behaves on the assumption that prices are unchanged and that unlimited quantities of demand and supply can be realized at the current prices. This conjecture is justified by the very fact that no exchange transactions are made and no trade contracts are in effect during the tâtonnement, until general equilibrium is established where prices are no longer changed, and every trader can purchase and sell exactly the amount he wishes at going prices.

In a monetary economy of the real world, where of course the tâtonnement assumption cannot be made and some exchange transactions actually take place before general equilibrium is established, even a competitive trader without power to control prices has to expect price changes and to try to sell when the price is high and to buy when the price is low, though he may not always succeed in doing so. This leads to the separation of sales and purchases, a separation which is made possible only by the use of money as the medium of exchange and the store of value. In Walrasian non-monetary real models where the tâtonnement assumption is made, on the other hand, sales and purchases are synchronized when general equilibrium is established so that there is no need for money, and indeed there is no reason why the role of medium of exchange should be exclusively assigned to a single item called money. Since equilibrium prices are already fixed and unchanged almost any non-perishable good can be used if necessary as a medium of exchange.

Walras considered tâtonnement even in his final model, that is, that of circulation and money. Since disequilibrium transactions are thus excluded and there is no uncertainty, there is no room here for money as a store of value. We have to assume therefore that people demand money only for the sake of convenience in transactions. Since all actual transactions are carried out at general equilibrium after the preliminary tâtonnement is over, however, this rationale for the demand for money is not at all convincing. The only role left for money is to determine its own price, that is, the general level of prices.

Walras gave two solutions for general equilibrium of each of his nonmonetary real models, as well as his monetary model. The first solution is the demonstration that the number of unknowns is equal to the number of independent equations, which Walras called the scientific or mathematical solution. But how can we find a solution of such equations, particularly when the number of equations is very large? The second solution of general equilibrium given by Walras (1926, pp. 162–3, 170–72) is tâtonnement itself, which is suggested by the mechanism of free competition in markets and is called the practical or empirical solution. Taking the example of the simple model of exchange these two solutions may be reformulated in modern notation as follows.

Consider an exchange economy of m goods and denote the price of and the excess demand for the jth good by p_j and E_j respectively. One condition for general equilibrium is that demand is equal to supply in all markets, that is

$$E_j(p_1,\ldots,p_m)=0,\quad j=1,\ldots,m. \tag{1}$$

In view of Walras's Law that

$$\sum_j p_j E_j \equiv 0, \tag{2}$$

only $(m-1)$ equations of (1) are independent, while we can assign the role of numéraire to the mth good so that $p_m=1$ since only relative prices are relevant in a non-monetary economy. Therefore (1) is replaced by

$$E_j(p_1,\ldots,p_{m-1})=0,\quad j=1,\ldots,m-1. \tag{3}$$

Equations 1 or 3 are derived from the competitive behaviour of individual consumer-traders. The ith consumer-trader is assumed to maximize his utility $U_i(x_{il},\ \ldots\ ,x_{im})$, subject to the budget constraint

$$\sum_j p_j x_{ij}=\sum_j p_j y_{ij} \tag{4}$$

where x_{ij} and y_{ij} denote respectively the gross demand for the jth good by the ith consumer-trader and the given initial holding of the jth good of the ith consumer-trader. The excess demand for the jth good is then defined as

$$E_j=\sum_i x_{ij}-\sum_i y_{ij}. \tag{5}$$

It is to be noted that excess demand is not defined in (1) and (3) explicitly as a function of the y_{ij}'s. The reason is that the y_{ij}'s are given constants and are assumed not to change through the process of exchange until the demand plans of all consumer-traders are simultaneously realized when general equilibrium is established. In other words the assumption of tâtonnement is already implicitly made in the mathematical or theoretical solution of general equilibrium.

The original form of Walrasian tâtonnement is the process of successive adjustment in each single market. Suppose the initial set of prices cried by the auctioneer $(p_1,\ \ldots\ ,p_{m-1})$ does not satisfy the condition (3) of general equilibrium, and we are for example in a situation described by

$$\begin{aligned} E_1(p_1,\ldots,p_{m-1})&>0\\ E_2(p_1,\ldots,p_{m-1})&<0\\ E_{m-1}(p_1,\ldots,p_{m-1})&>0 \end{aligned} \tag{6}$$

The price of the first good p_1 is now adjusted by reference to its excess demand E, and increased in the situation (6) until an equilibrium in the first market is established, that is

$$E_1\left(p_1',p_2,\ldots,p_{m-1}\right)=0. \tag{7}$$

Here E_1 is assumed to be decreasing with respect to p_1, an assumption which, writing the partial derivative of the excess demand function for the jth good with respect to the kth price by E_{jk}, may be symbolized by $E_{11}<0$.

Under the new price system $\left(p_1',p_2,\ldots,p_{m-1}\right)$ the remaining $m-1$ markets may or may not be in equilibrium. If the second market is out of equilibrium, again under the assumption that $E_{22}<0$, the price of the second good is changed from p_2 to p_2' so as to satisfy

$$E_2\left(p_1',p_2',p_3\ldots,p_{m-1}\right)=0. \tag{8}$$

(Generally, this will upset the equilibrium in the first market (7)). Under the price system $\left(p_1',p_2',p_3\ldots,p_{m-1}\right)$, then, the price of the third good p_3 is adjusted if the third market (where $E_{33}<0$) is out of equilibrium, upsetting the equilibrium in the second market (8) just established. In this way the last, $m-1$th market, where $E_{m-1,\ m-1}<0$, is eventually cleared by changing the price system from $\left(p_1',\ldots p_{m-2}',p_{m-1}\right)$ into $\left(p_1',\ldots p_{m-2}',p_{m-1}'\right)$ so as to satisfy

$$E_{m-1}\left(p_1',\ldots p_{m-2}',p_{m-1}'\right)=0. \tag{9}$$

By this time all the markets except the last, which were once cleared successively, have generally been thrown out of their respective equilibria again. Neither the price system we have just arrived at, $\left(p_1',\ldots,p_{m-1}'\right)$, nor the initial system, $(p_1,\ \ldots\ ,p_{m-1})$, is part of a general equilibrium.

The question then is which of the systems is closer to a true general equilibrium that satisfies (3). Walras argued that the former price system is closer to equilibrium than the latter since for example $E_1(p'_1, \ldots p'_{m-1}) \neq 0$ is closer to 0 than $E_1(p_1, \ldots p_{m-1}) \neq 0$. The reason for this, according to Walras, is that the change from p_1 to p'_1 which established (7) exerted a direct influence that was invariably in the direction of zero excess demand so far as the first good is concerned. But the subsequent changes from p_2 to $p'_2, \ldots, p_{m-1}$, to p'_{m-1}, which jointly moved the first excess demand away from zero, exerted only indirect influences, some in the direction of equilibrium and some in the opposite direction, at least so far as the excess demand for the first good is concerned. So up to a certain point they cancelled each other out. Hence, Walras concluded, by repeating the successive adjustment of $m-1$ markets along the same lines, that is, changing prices according to the law of supply and demand, we can move closer and closer to general equilibrium.

Walras's argument for the convergence of the tâtonnement process to general equilibrium was intended to be, if successful, the first demonstration of the existence of competitive general equilibrium (Wald 1936). As we said above, it was merely an argument for the plausibility of such convergence of the process of tâtonnement, and cannot be considered as a rigorous demonstration of existence of equilibrium. Whether indirect influences of the prices of other goods on the excess demand of a given good cancel each other out will certainly depend on substitutability and complementarity between goods. For example, indirect influences are *not* cancelled out and the excess demand of a good *is* increased if the prices of all gross substitutes are raised and the prices of all gross complements are lowered. In addition to the Walrasian stability condition for a single market, that is, $E_{jj} < 0$ for all j, therefore, some conditions on the cross-effects of prices on excess demands, that is on E_{jk}, $j \neq k$, have to be imposed so as to demonstrate convergence.

It was Allais (1943, vol. 2, pp. 486–9) who first demonstrated the convergence of Walrasian tâtonnement by assuming gross substitutability, that is, $E_{jk} > 0$ for all $j \neq k$. To see whether the price system moves closer and closer to the general equilibrium, which he assumes to be at least locally unique, Allais defines the distance D of a price system from the equilibrium price system as the sum of the absolute values of the value of excess demand for all goods, including the numéraire. The convergence of tâtonnement is then demonstrated by showing that this distance D is always decreased by changes in prices that are made in accordance with the law of supply and demand. His demonstration may be reformulated in our notation as follows.

The distance to the general equilibrium is defined as

$$D = \sum_j |p_j E_j| \tag{10}$$

where the summation runs from $j = 1$ to $j = m$, and E_j is defined as a function of $p_1, \ldots, p_{m-1}$ as in (3). In view of Walras' Law (2), D can be replaced either by the summation of positive excess demands

$$D_1 = \sum_j p_j \max(0, E_j) \tag{11}$$

or by the summation of negative excess demands

$$D_2 = -\sum_j p_j \min(0, E_j) \tag{12}$$

where $\max(0, E_j)$ denotes E_j if it is positive and 0 if E_j is negative, and $\min(0, E_j)$ denotes E_j if it is negative and 0 if E_j is positive. From (2), that is $D_1 - D_2 = 0$, it is clear that

$$D = 2D_1 = 2D_2 \tag{13}$$

so that whether D is increasing or decreasing can be seen by checking whether D_1 or D_2 (whichever is more convenient) is increasing or decreasing.

Suppose E_1 to be positive as in (6) and that p_1 is raised following the law of supply and demand. From (12), we have

$$\partial D_2 / \partial p_1 < 0 \tag{14}$$

since $E_{j1} > 0$ for any j such that $E_j < 0$, from gross substitutability. In other words, a change in the price of the first good from p_1 to p_1' so as to satisfy (7) decreases the sum of negative excess demands D_2 and therefore the distance D to the general equilibrium. Suppose next that $E_2(p_1', p_2, \ldots, p_{m-1})$, is negative and p_2 is lowered to p_2' so as to satisfy (8). From (11) this time, we have

$$\partial D_1 / \partial p_2 > 0 \tag{15}$$

since $E_{j2} > 0$ for any j such that $E_j > 0$ from gross substitutability. In other words, a decrease in the price of the second good from p_2 to p_2' decreases the sum of positive excess demands D_1 and therefore the distance D to the general equilibrium.

Generally, if E_j is positive and p_j is raised D is decreased, which can be seen from the fact that D_2 is decreased. Similarly, if E_j is negative and p_j is lowered again D is decreased, which can be seen from the consideration of the behaviour of D_1. Out of the general equilibrium D remains positive and there exists at least one non-numéraire good with non-zero excess demand, so that its price is changing. The distance to the general equilibrium always decreases out of equilibrium, and therefore we can move closer and closer to that equilibrium by changing prices according to the law of supply and demand, provided that gross substitutability is assumed.

Though Walras discussed the behaviour of the process of successive adjustment, he was not against the consideration of *simultaneous adjustment processes* in all markets (Uzawa 1960; Jaffé 1981). If we assume that adjustments take place not only simultaneously but also continuously, the tâtonnement process that each rate of change of price is governed by excess demand can be described by a set of differential equations,

$$\mathrm{d}p_j/\mathrm{d}t = a_j E_j(p_1, \ldots, p_{m-1}),\ j = 1, \ldots, m-1, \tag{16}$$

where t denotes time and the a_j's are positive constants signifying the speed of adjustment in the jth market. The study of the behaviour of the solutions of (16), that is prices as functions of t, which was initiated by Samuelson (1941) is called the study of the *stability of competitive equilibrium* and has been extensively carried out by many mathematical economists (Arrow and Hahn 1971, pp. 263–323; Negishi 1972, pp. 191–206). It is well known that gross substitutability is also a sufficient condition for the *convergence of adjustment processes* like (16).

The idea of tâtonnement was clearly suggested to Walras from the observation of how business is done in some well organized markets in the real world, like the stock exchanges, commercial markets, grain markets, fish markets. As a matter of fact, Walras was well informed of the actual operation of the Paris Stock Exchange where disequilibrium transactions actually did not occur (Jaffé 1981). Tâtonnement is therefore not entirely unrealistic as a model of adjustment in such special markets.

However, it is certainly very unrealistic to apply such a model of special markets to the whole economy, since preliminary adjustments are usually not made before exchange transactions and effective contracts take place, even in markets where competition, though not so well organized, functions fairly satisfactorily. Of course, Walras would have admitted this, since tâtonnement was for him not so much a description of the process of adjustment in the markets of the real world as it was the demonstration of the existence of general equilibrium, that is a limit to which tâtonnement converges. It should be so interpreted not only in the case of successive tâtonnement, which reminds us of the Gauss–Seidel method of solving a set of simultaneous equations, but also in the case of simultaneous tâtonnement (16), where time t is not real calendar time, but hypothetical process time. This is no wonder, since Walrasian non-monetary models are not intended to be faithful descriptions of the real world. They are designed rather to make clear the significance of each component of the market economy and to uncover the general principles that underlie its working.

One may feel that such an interpretation of Walrasian tâtonnement is too strict and that the behaviour of not so well organized markets can be described approximately by the tâtonnement model. Walrasian tâtonnement may be interpreted

as something like the laws of motion, that work strictly speaking only in the ideal frictionless world but which can be applied approximately to the real world. The law of supply and demand can certainly be applied even in markets where there is no auctioneer, traders are dispersed, and exchange transactions take place and effective contracts are made before equilibrium of demand and supply is established.

Prices are formed differently in each exchange transaction by negotiation between relevant parties of traders. The law of indifference tends to prevail, however, if the transmission of information is nearly perfect, since atomistic traders know the difficulty of purchasing (selling) at prices lower (higher) than the prices offered by competitors and there are, furthermore, arbitrage activities. If demand falls short of supply, it is suicidal for atomistic sellers to offer a price higher than the average market price, while an atomistic purchaser is unable to consider a price lower than the average market price when demand exceeds supply. With disequilibrium of supply and demand, exchange transactions can take place only if demanders (suppliers) can find suppliers (demanders). If demand is deficient therefore sellers consider cutting prices or increasing marketing costs in order to attract more purchasers, since a drastic increase in sales is expected from slight falls in price or slight increases in marketing costs when information is nearly perfect. By observing such behaviour by the sellers, the purchasers also insist on price cuts. Thus price falls in the face of excess supply. Similarly, market prices rise as the result of a similar process of disequilibrium exchange transactions in the face of excess demand, in which the roles of sellers and purchasers are interchanged from the case of excess supply.

Therefore, we can extend (16) to

$$\begin{aligned} \mathrm{d}p_j/\mathrm{d}t &= a_j E_j(p_1, \ldots, p_{m-1}, y_{11}, \ldots, y_{nm}), \\ j &= 1, \ldots, m-1, \end{aligned} \tag{17}$$

where the E_j's are again derived from (5) but have now to be considered explicitly as functions of the y_{ij}'s that is the stock of the jth good held by the ith consumer-trader, $i = 1, \ldots, n$, since the y_{ij}'s are no longer constants but instead are changed by disequilibrium transactions among the n consumer-traders. Here we cannot discuss in detail how the y_{ij}'s are changed as a result of transactions at disequilibria, and have to be content with the general assumption that their rates of change depend on everything, that is, we have

$$\begin{aligned} \mathrm{d}y_{ij}/\mathrm{d}t &= F_{ij}(p_1, \ldots, p_{m-1}, y_{11}, \ldots, y_{nm}), \\ i &= 1, \ldots, n \quad j = 1, \ldots, m, \end{aligned} \tag{18}$$

where the F_{ij}'s are unknown functions incorporating rules for exchange transactions out of equilibria. Models of an economy with (17) and (18) are called nontâtonnement models or *non-recontracting models*.

Generally, if a non-tâtonnement or non-recontracting process (17) and (18) converges, it does so to an equilibrium that is different from that arrived at by the tâtonnement process (16), since changes in the y_{ij}'s due to disequilibrium exchange transactions have effects on (17) which do not exist in the case of (16). As Newman (1965, p. 102) correctly pointed out, however, the difference can be safely neglected, if the speed of price adjustment in every market is very high (that is, the a_j's in (17) are very large), since then markets arrive at equilibrium prices so rapidly that the effects of disequilibrium transactions are prevented from becoming serious. Although the possibility of disequilibrium transactions is not institutionally excluded and there may well be some, most transactions are actually carried out at equilibrium so that it looks as if the assumption of tâtonnement is satisfied. In this sense, tâtonnement models can be used to describe the behaviour of non-tâtonnement or non-recontracting markets in the real world.

Although the tâtonnement model can be applied to markets that are not so well organized if the transmission of information is nearly perfect and the speed of price adjustment is rapid, the general equilibrium tâtonnement model (16) is still not a realistic description of the real world economy. The reason is that the role of money as

the medium of exchange and a store of value is very important in the real world, while as we saw it is highly limited in a model where most exchange transactions are simultaneously carried out at equilibrium. To make our model more realistic so that sales and purchases take place at disequilibria and are separated by the use of money, therefore, we have to get rid of tâtonnement by arguing that the speed of price adjustment is not rapid in (17), so that disequilibrium transactions cannot be ignored.

If the transmission of information is perfect, the law of supply and demand can be applied even in not so well organized markets where no auctioneer exists and disequilibrium transactions take place. This is because every seller (purchaser) perceives an infinitely elastic demand (supply) curve and expects that a drastic increase in sale (purchase) is made possible by a slight reduction (increase) in price. If total demand falls short of total supply in a market, then every trader willingly reduces price or accepts a reduction in it. Similarly, if total demand exceeds total supply in a market every trader willingly raises price or accepts a rise in it.

The transmission of information may not be so perfect, however, in markets where traders are dispersed and so cannot meet in a big hall as they do in the case of Walrasian tâtonnement. Suppose that a market is segmented and transmission of information is perfect among closely related traders, but that it is not so between different segments. Individual traders are assumed to keep contact with current trade partners and not to leave the segment of the market in which they are currently located in search of more favourable trade conditions, unless either they are well informed of such conditions in other segments or trade conditions change unfavourably in the original segment. Possibly because of consideration of cost, traders are constrained by inertia and do not move unless shocked by information on other segments or by changes in the original segment.

Then even an atomistic seller (purchaser) does not perceive an infinitely elastic demand (supply) curve. A seller expects that sales cannot be increased very much by reduction of price since only those purchasers who are currently buying from him are well informed of the price reductions, and this information is not perfectly transmitted to those purchasers who belong to other segments of the market and who are not buying from him. When total demand falls short of total supply and other sellers do not raise the price, it cannot be expected that 'their' purchasers leave them in search of cheaper sellers. The same seller has to expect, however, that sales will be drastically reduced if the price is raised, since those customers who are currently buying will be well informed of this and will leave to search for cheaper sellers, which they can find easily when total demand falls short of total supply and there are many other sellers willing to sell more at the unchanged price.

Atomistic sellers perceive kinked demand curves, with a downward sloping segment for levels of sale higher than the current one, and an almost infinitely elastic segment for levels of sale lower than the current one, when the market is in excess supply. It is very likely then that price does not fall and remains sticky in the face of excess supply (Reid 1981, pp. 65–6, 96–9; Negishi 1979, p. 36). It may not pay to reduce price if demand cannot be increased very much. Similarly, an atomistic purchaser perceives a kinked supply curve with an upward sloping segment for levels of purchase higher than the current one, and an almost infinitely elastic segment for levels of purchase lower than the current one, when the market is in excess demand. Since the transmission of information is imperfect and the purchaser cannot attract many sellers by raising price, it may not pay to raise price even if a larger purchase is wanted at the current price. It is very likely, therefore, that price does not rise and remains sticky in the face of excess demand.

Thus prices may be sticky and may not be adjusted quickly by demand and supply in not-well-organized markets in the real world. The speed of adjustment in (17) need not be rapid enough to allow one to ignore the effects of disequilibrium transactions, so that the tâtonnement process (16) cannot then be regarded as a realistic description of adjustment in real-world markets.

Walrasian tâtonnement models are, of course, not designed to describe such markets empirically. They are constructed to show how the market mechanism works beautifully under ideal conditions. No one can deny that Walrasian economics succeeded in accomplishing this purpose. The market mechanism, however, does not work so beautifully in the real world. It certainly manages to work somehow but quite often at the cost of prolonged disequilibria in markets, such as involuntary unemployment in the labour market and excess capacity in goods markets. This is why we have to supplement Walrasian economics by launching out into the study of non-Walrasian economics.

Recontracting. Since the idea of recontracting is due originally to Edgeworth, who developed it in a way different from Walras's tâtonnement, the implication of Edgeworth's theory of recontract has to be carefully considered in its relation to the theory of tâtonnement in Walrasian economics. These two theories are different from each other in at least two ways, namely with respect to the law of indifference (the uniformity of prices) and to the provisional nature of revocability of trade contracts. The first problem is discussed below, while the second will be considered in the next section.

There have been different interpretations as to whether Edgeworth's *Mathematical Psychics* (1881) excluded disequilibrium transactions or assumed the irrevocability of contracts (Walker 1973; Creedy 1980). Even if we assume that disequilibrium transactions are excluded, however, the theory of recontract in Edgeworth is different from the theory of tâtonnement in Walrasian economics. The law of indifference (that is, the existence of uniform market prices even in disequilibria) is imposed as an axiom in the original Walrasian as well as in modern Walrasian economic theories. This axiom may be justified either through arbitrage activities or by the existence of the auctioneer, and enables individual traders to act as price takers who have only to adjust their plans of supply and demand to the given prices. Such an axiom is not imposed in Edgeworth's recontracting model.

To demonstrate his famous limit theorem (Bewley 1973), Edgeworth starts with a simple two-good two-individual model of exchange, where a trader X offers a good x to a trader Y in exchange for a good y. If we consider the so-called Edgeworth box diagram, any point on the contract curve, where each of two individual traders is not worse off than before exchange, can be a final settlement of trade contract which cannot be varied by recontract. To narrow down the range of possible final settlements Edgeworth introduces a second X and a second Y, each respectively identical to the first, both in tastes and initial endowments. Since identical traders have to be treated equally in any final settlement, we can still use the same box diagram. Now it can be shown that no final settlement of contract can contain points on the contract curve which give 'small' gains from trade to the X traders. Otherwise, it is 'possible for one of the Ys (without the consent of the other) to *recontract* with the two Xs, so that for all those three parties the recontract is more advantageous than the previously existing contract' (Edgeworth 1881, p. 35). Similarly, it is possible to exclude as final settlements those points which give 'small' trade gains to Y traders.

In this way Edgeworth shows that the range of possible final settlements shrinks as the number of identical traders grows. If there are infinitely many traders the only remaining final settlements turn out to be precisely the points of Walrasian equilibrium, each with a uniform price line, that is the common tangent to indifference curves of X and Y passing through the point of initial endowments. In the terminology of the modern theory of cooperative games, the core of the exchange game (that is, those allocations not blocked by any coalitions of players) consists only of the Walrasian equilibria when the numbers of the Xs and the Ys are each infinitely large. Thus Edgeworth tries to show that the recontracting process in the large economy, where traders obtain a free flow of information through the making and breaking of provisional contracts, leads to the same uniform prices that are given by the auctioneer to price-taking traders in Walrasian equilibria. Though there are no uniform market prices and individual

traders are not assumed to be price-takers in Edgeworth's recontracting process, the resulting equilibrium exchanges are the same as those obtained through Walrasian tâtonnement in a large economy. In such an economy, therefore, where information is perfect, we can safely argue as if there were uniform market prices and as if traders were price-takers. In a sense, Edgeworth justified the Walrasian axiom, since axioms of theories should be assessed not by themselves but by the results derived from them. Even if the Walrasian axiom is not itself realistic, the results derived from it can be as realistic as those derived from more realistic but more complicated axioms.

In later writings Edgeworth confirmed his early position on Walras and the uniformity of prices. Walras

> describes *a* way rather than *the* way by which economic equilibrium is reached . . . Walras's laboured description of prices set up or 'cried' in the market is calculated to divert attention from a sort of higgling which may be regarded as more fundamental than his conception, the process of *recontract* . . . The proposition that there is only one price in a perfect market may be regarded as *deducible* from the more axiomatic principle of recontract. (*Mathematical Psychics*, p. 40 and context: Edgeworth 1925, vol. 2, pp. 311–23)
>
> We may add that even the existence of a uniform rate-of-exchange between any two commodities is perhaps not so much axiomatic as deducible from the process of competition in a perfect market. (Edgeworth 1925, vol. 2, p. 453)

It is possible to interpret Edgeworth's theory of bilateral exchange (Edgeworth 1925, vol. 2, pp. 316–19) as a theory of a process where not only the rate of exchange is variable but also contracts are irrevocable. Starting from a situation with initial holdings, two goods are actually exchanged so as always to increase the utility of each of the two traders. Since exchanges are irrevocable, however, where on the contract curve this process of exchange will terminate depends on the path of exchanges as well as on the initial holdings. Hence it contrasts strongly with Walrasian tâtonnement, the equilibrium of which depends only on the initial holdings. There is no confusion, however, between this theory of Edgeworth and Edgeworth's theory of recontract interpreted in the sense of tâtonnement, since the modern extension of the former theory to the case of multiple traders is rightly called the theory of Edgeworth's *barter* process (Uzawa 1962; Fisher 1983, pp. 29–31).

Incidentally, Edgeworth's idea that exchanges necessarily take place only in the direction of increasing utilities can be relevant only in a barter economy. In a monetary economy an exchange of one good against another is decomposed into an exchange of the first good against money and an exchange of money against the second good. Even though the completed exchange of the two goods increases utility, its first half need not do so since in the course of the exchange process one may temporarily receive more money than one plans to keep eventually. In other words, one may impose a rule for non-monetary goods of no overfulfilment of demand and supply plans in the process of exchange, but this cannot be done for money, which has to act both as the medium of exchange and as a store of value beyond the current period.

In view of the current usage of the concept of recontracting in the sense of tâtonnement, what is confusing is the fact that Edgeworth sometimes, and particularly in his later writings, applied his recontracting model to situations where exchange transactions actually take place at disequilibria. To show that his model is of more than academic interest Edgeworth considered the case of a labour market, which each day ends in disequilibrium after exchange transactions have taken place at disequilibrium rates of exchange. From day to day, as the traders' knowledge of the state of the disequilibrium in the market changes they progressively modify their behaviour, changing the rate of exchange in such a way that the market converges to equilibrium.

Since labour service is perishable within a day and the number and dispositions of the traders are assumed to be unchanged, this process over a sequence of days is formally equivalent to the recontracting process within a day, even though in the former process contracts made on the previous days are irrevocable while in the latter

disequilibrium contracts are revocable. Edgeworth insisted that in this example of a process over a sequence of days (Edgeworth 1925, vol. 1, p. 40) traders do recontract, in the full sense of *Mathematical Psychics*. Since contracts made in earlier days are irrevocable, however, in this case to recontract implies that a new contract is made which is different from that carried out on the previous day. It does not imply the cancellation of contracts already made.

Only a formal similarity exists between these two processes of recontracting, which is due to the assumption that disequilibrium exchange transactions do not really involve a permanent redistribution of wealth. Although labour service is perishable within a day, however, the money paid against labour service certainly is not and it is likely that a redistribution of wealth does take place over a sequence of days. Even from a formal point of view, then, Edgeworth's model of the labour market is rather a pioneering instance of *non-recontracting models.*

No one can deny that the rigorous demonstration of the dynamic stability of tâtonnement under certain sufficient conditions has substantially improved on the original argument for the plausibility of its convergence that was made by Walras. More importantly, however, the recent studies on stability have helped us to understand the underlying economic assumptions of the Walrasian tâtonnement process itself, and made us realize its considerable differences from most price adjustment processes in actual economies. The similar studies of Edgeworth's recontracting process have been helpful in the same way.

As we have shown, Walrasian tâtonnement is a realistic approximation to some actual adjustment processes, provided that the transmission of information is perfect and the speed of adjustment is rapid, as is roughly the case in well organized markets. The problem that remains to be studied, therefore, is the nature of adjustment processes when these conditions are not satisfied, that is, when markets are not so well organized. This is the problem of non-recontracting models in non-Walrasian or disequilibrium economies.

See Also

- ▶ Auctioneer
- ▶ General Equilibrium
- ▶ Walras, Léon (1834–1910)

Bibliography

Allais, M. 1943. *Traité d'économie pure*. 2nd ed. Paris: Imprimerie Nationale. 1952.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Bewley, T.F. 1973. Edgeworth's conjecture. *Econometrica* 41: 425–454.

Creedy, J. 1980. Some recent interpretations of mathematical psychics. *History of Political Economy* 12: 267–276.

Edgeworth, F.Y. 1881. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. London: C. Kegan Paul & Co..

Edgeworth, F.Y. 1925. *Papers relating to political economy*. Vol. 3. London: Macmillan.

Fisher, F.M. 1983. *Disequilibrium foundations of equilibrium economics. Econometric society monographs in pure theory*. Cambridge: Cambridge University Press.

Hicks, J.R. 1934. Léon Walras. *Econometrica* 2: 338–348.

Jaffé, W. 1967. Walras's theory of *tâtonnement*: A critique of recent interpretations. *Journal of Political Economy* 75: 1–19.

Jaffé, W. 1981. Another look at Léon Walras's theory of *tâtonnement*. *History of Political Economy* 13: 313–336.

Kaldor, N. 1934. A classificatory note on the determinateness of equilibrium. *Review of Economic Studies* 1 (February): 122–136.

Marshall, A. 1890. *Principles of economics*. London: Macmillan.

Morishima, M. 1977. *Walras' economics: A pure theory of capital and money*. Cambridge: Cambridge University Press.

Negishi, T. 1972. *General equilibrium theory and international trade*. Amsterdam: North-Holland.

Negishi, T. 1979. *Microeconomic foundations of Keynesian macroeconomics*. Amsterdam: North-Holland.

Newman, P. 1965. *The theory of exchange*. Englewood Cliffs: Prentice-Hall.

Patinkin, D. 1956. *Money, interest, and prices: An integration of monetary theory*. Evanston: Row, Peterson.

Reid, G.C. 1981. *The kinked demand curve analysis of oligopoly: Theory of evidence*. Edinburgh: Edinburgh University Press.

Samuelson, P.A. 1941. The stability of equilibrium: Comparative statics and dynamics. *Econometrica* 9: 97–120.

Uzawa, H. 1960. Walras' tâtonnement in the theory of exchange. *Review of Economic Studies* 27: 182–194.

Uzawa, H. 1962. On the stability of Edgeworth's barter process. *International Economic Review* 3: 218–232.
Wald, A. 1936. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie* 7: 637–670.
Walker, D.A. 1973. Edgeworth's theory of recontract. *Economic Journal* 83: 138–149.
Walras, L. 1874–1877. *Éléments d'économie politique pure ou théorie de la richesse sociale.* Definitive edn, Lausanne, 1926. Trans. W. Jaffé as *Elements of Pure Economics*, London: George Allen and Unwin; Homewood, IL: Richard D. Irwin, 1954.

Taussig, Frank William (1859–1940)

Warren J. Samuels

Keywords

Clark, J. B.; Comparative advantage; Economic sociology; Marginal productivity theory; Marginalism; Mathematics and economics; Minimum wages; Monopoly; Protection; Schumpeter, J. A.; Specie-flow mechanism; Taussig, F. W.; Trade unions; Unemployment insurance; Wages fund

JEL Classifications

B31

Taussig was born on 28 December 1859 in St Louis, Missouri, and died on 11 November 1940 in Cambridge, Massachusetts. After starting college at Washington University, St Louis, he transferred to Harvard University, where he received the BA (1879), Ph.D. (1883) and LLB (1886). He also studied at the University of Berlin.

Taussig was one of the foremost US economists for half a century. He was on the Harvard faculty from 1885 to 1935, where he was a magisterial teacher and edited the *Quarterly Journal of Economics* from 1896 to 1936. A member of several government commissions, he was the first chairman of the US Tariff Commission, 1917–19, and an adviser to President Woodrow Wilson. He was president of the American Economic Association in 1904 and 1905. His *Principles of Economics* (1911) was the foremost US textbook for generations of economists and noneconomists.

Taussig was accurately called 'the American Marshall' by Joseph Schumpeter because of his professional stature. He shared with Alfred Marshall an identification with the Ricardo–Mill tradition coupled with a willingness to integrate the ideas of marginalism; a scepticism of the mathematicization of economics; a desire to moderate conflict within the discipline; an understanding that economics was or should be more an organon of analysis, a collection of tools, than a body of doctrine; a sympathy for the working class; and a view that economics was to remain political economy, to include what Schumpeter called 'economic sociology'. He preferred J.S. Mill's *Principles of Political Economy* to any modern text, including Marshall's *Principles,* because in his view it prevented delusions as to economic questions being easy of solution. Like Marshall, too, he considered the Austrian system needlessly complex.

His principal work as an economic theorist lay in wage theory and in international trade theory and policy. In the former, he attempted to resuscitate a modified version of the wages fund theory, centring on the relative inelasticity of the short-run supply of consumer goods, which he combined with marginal productivity theory. He stressed, however the role of non-competing groups (as part of his great realism in matters of stratification), criticized John Bates Clark's moralistic version of marginal productivity theory and argued that the frequent superior advantage in bargaining of employers meant that marginal productivity was only an upper limit, in the absence of effective competition.

In the field of international trade, in which he was the principal US figure for decades, his major concerns were the complexities of comparative advantage, the role of the specie flow mechanism and the international trade mechanism under non-specie monetary systems, and the history and analysis of protection. His position on protectionism was complex: he affirmed free trade, accepted the infant-industry argument with considerable scepticism of its application in practice, favoured

gradual lessening of tariffs and (or but) affirmed a stable tariff system rather than policies of disruptive shifts and shocks.

In other policy controversies he strenuously opposed the free coinage of silver as inflationary; criticized unemployment insurance and minimum wages as violating traditional individual initiative; thought progressive taxation less important than the elimination of monopoly and the use of education to diffuse opportunity; and, understanding of worker reliance on unions, blamed labour–management conflict on the failure of management to exercise the responsibilities of wealth and power and to be more understanding of worker interests and ambitions.

Taussig's political economy, or economic sociology, set him off from most other leading orthodox economists. He saw society as both a structure and a struggle for power and privilege, in which an instinct of domination and an impulse of emulation were prevalent if not dominant. These ideas pervaded his *Principles*, in which he presented, for example, a functional analysis of the leisure class and, in his *Inventors and Money Makers* (1915b) and (with C.S. Joslyn) *American Business Leaders* (1932), a further analysis of both the complex psychological bases of economic behaviour and the role of leadership in the successful operation of the market mechanism. Schumpeter wrote that Taussig was 'among those few economists who realize that the method by which a society chooses its leaders, in what for its particular structure, is the fundamental social function ... is one of the most important things about a society, most important for its performance as well as for its fate' (Schumpeter 1951, p. 217). In these respects, Taussig's work was compatible with institutionalism, but his establishment position apparently kept those in the later tradition from fully appreciating his contributions. Taussig's ideas here, supplementing those of Friedrich von Wieser, had some influence on Schumpeter himself.

See Also

► Wages Fund

Selected Works

1888. *The tariff history of the United States,* 8th ed, revised. New York: Putnam's, 1931.
1896. *Wages and capital.* New York: D. Appleton.
1911. *Principles of economics,* 4th ed. New York: Macmillan, 1939.
1915a. *Some aspects of the tariff question,* 3rd ed, enlarged. Cambridge, MA: Harvard University Press, 1931.
1915b. *Inventors and money makers.* New York: Macmillan.
1927. *International trade.* New York: Macmillan.
1932. (With C.S. Joslyn.) *American business leaders: A study in social origins and social stratification*. New York: Macmillan.

Bibliography

Opie, R. 1941. Frank William Taussig (1959–1940). *Economic Journal* 51: 347–368.
Schumpeter, J.A. 1951. *Ten great economists*. New York: Oxford University Press.

Tawney, Richard Henry (1880–1962)

J. M. Winter

Keywords
Capitalism; Individualism; Protestant ethic; Religion and economic development; Tawney, R. H.; Weber, M.

JEL Classifications
B31

R.H. Tawney was an economic historian and socialist philosopher whose Anglican beliefs lay at the heart of his influential studies of the enduring problem of the ethics of wealth distribution. As Professor of Economic History at the London School of Economics from 1921 to 1958, he became the doyen of a school of thought which

defined the subject as the exploration of the resistance of groups and individuals in the past to the imposition on them of capitalist modes of thought and behaviour.

In his first book, *The Agrarian Problem in the Sixteenth Century* (1912) – written to provide an appropriate text for his pioneering tutorial classes for the Workers' Educational Association – he examined patterns of rural development, protest and litigation surrounding the enclosure of land in Tudor England. After service in the British Army during the First World War – he was severely wounded on the first day of the Battle of the Somme – Tawney returned to his scholarship and developed the arguments which appeared in perhaps his best-known work, *Religion and the Rise of Capitalism* (1926). Here he showed how alien to the teachings of the Reformation was the assumption that religious thought had no bearing on economic behaviour. Tawney captured in classical prose the clash within religious opinion that preceded that abnegation of the social responsibility of the churches and suggested that 'religious indifferentism' was but a phase in the history of Christian thought.

In *Religion and the Rise of Capitalism* Tawney crystallized a number of ideas he had begun to consider in the pre-1914 period. In a commonplace book he kept from 1912 to 1914, Tawney jotted down notes on many of his religious and historical preoccupations. Among them is the simple query, 'I wonder if Puritanism produced any special attitude toward economic matters'. Over the following decade, he gathered evidence on this subject, and presented preliminary statements in the Scott Holland Memorial Lectures at King's College, London, in 1922, and in the lengthy introduction he wrote for a 1925 edition of Thomas Wilson's *Discourse on Usury* of 1569.

In the notes Tawney left concerning this facet of his historical research, there is no evidence whatsoever that he drew on the celebrated essay of Max Weber, originally published in 1905, on *The Protestant Ethic and the Spirit of Capitalism*. Indeed, a full appreciation of Tawney's Anglican concerns requires a divorce between the two partners of the so-called 'Tawney–Weber' thesis.

It is true that both men believed that (in Tawney's words) 'The fundamental question to be asked, after all, is not what kind of rules a faith enjoins, but what kind of character it values and cultivates.' They agreed as well that there was in Calvinism a corrosive force which undermined traditional doctrines of social morality in ways which would have shocked the early reformers. And they shared the view that in Protestant teaching there was an important emphasis in religious terms on the 'inner isolation of the individual' which reinforced a more general individualism of social and economic behaviour.

But what differentiates their work is the uses to which they put their interpretations of Protestantism. Weber's essay was but one part of a comprehensive study of the sociology of religion. It reflects his overriding concern with the development of what he termed the rational bureaucratic character of modern society. In both these facets of his work he charted the progressive, relentless, and irreversible demystification of the world.

Weber's essay helped foster a belief in the bleak permanence of the spirit of capitalism which Tawney laboured to refute throughout his work. *Religion and the Rise of Capitalism* was written precisely to counter the view that social indifferentism in religious thought and individualism in economic thought were unchangeable features of modern life. If Weber's purpose was to describe the demystification of the world, Tawney's was to help in the demystification of capitalism, by stripping it of some of its most powerful ideological supports, derived from one reading of the Protestant tradition.

Anglicanism is, of course, a house of many mansions, in which there is room for reactionaries and socialists alike. The view that capitalism is unchristian because it stultifies the common fellowship of men of different means and occupations has never been more than a minority view. But, at precisely the same time as he was writing *Religion and the Rise of Capitalism*, Tawney joined a number of other influential Anglicans who spoke out against capitalism as a way of life which violated the moral precepts of their faith.

This position was as evident in his essays in political philosophy as it was in his scholarship in

economic history. In *The Acquisitive Society* (1921) and in *Equality* (1931), Tawney argued that capitalism was an irreligious system of individual and collective behaviour, since it was based on the institutionalization of distinctions between men based on inherited or acquired wealth. For a Christian, such divisions manifested a denial of the truth that all men are equally children of sin and equally insignificant in the eyes of the Lord. What Matthew Arnold had called the 'religion of inequality' was really the obverse of a Christian way of looking at the world.

Tawney's legacy has been particularly pervasive, because his voice had a resonance which appealed to many who did not share his religious outlook. This was in part because he wrote with the moral outrage of Marx and with the grace and eloquence of Milton. His strength lay too in the fact that his was a distinctively English voice. This did not prevent his advocacy of the comparative method in the study of economic history, best evidenced in his book *Land and Labour in China* (1932), written after an eight-month mission to China as an educational adviser to the League of Nations, and in a history of the American labour movement he wrote while adviser to Lord Halifax, British Ambassador to Washington during the Second World War.

But Tawney's influence lies more centrally in his writings on the moral issues posed by capitalist economic development in Britain. His call for an alternative to the cash nexus – firmly within the tradition of Owen, Ruskin and Morris – has continued to strike a chord among many people not of religious temperament who have sought indigenous answers to the problems of a society crippled by the injuries of class.

See Also

- ▶ Fabian Economics
- ▶ Weber, Max (1864–1920)

Selected Works

1912. *The Agrarian problem in the sixteenth century.* London: Longmans.
1914. (ed., with A.E. Bland and P.A. Brown.) *English economic history: Selected documents.* London: Bell.
1921. *The acquisitive society.* London: Bell.
1924. (ed., with E. Power.) *Tudor economic documents.* London: Longmans.
1926. *Religion and the rise of capitalism.* London: Murray.
1927. ed. *Economic history: The collected papers of George Unwin.* London: Macmillan.
1931. *Equality.* London: Allen & Unwin.
1932. *Land and labour in China.* London: Allen & Unwin.
1953. *The attack and other essays.* London: Allen & Unwin.
1958. *Business and politics under James I: Lionel Cranfield as merchant and minister.* Cambridge: Cambridge University Press.
1964. *The radical tradition.* London: Allen & Unwin.
1972. *R.H. Tawney's commonplace book,* ed. D.M. Joslin and J.M. Winter. Cambridge: Cambridge University Press.
1978. *History and society. Essays by R.H. Tawney,* ed. J.M. Winter. London: Routledge & Kegan Paul.
1979. *The American labour movement and other essays,* ed. J.M. Winter. Brighton: Harvester Press.

Bibliography

Terrill, R. 1974. *R.H. Tawney and his times*. Cambridge, MA: Harvard University Press.
Winter, J.M. 1974. *Socialism and the challenge of war.* London: Routledge & Kegan Paul.

Tax Competition

Michael Keen

Abstract

Tax competition refers to strategic tax-setting in a non-cooperative game between

jurisdictions – whether countries or states or provinces within a federation – with each setting some parameters of its tax system in relation to the taxes set by others. This creates a potential case for international tax coordination, though the revenue impact has as yet been modest (at least in OECD countries). Conflicting national interests make it difficult to design effective coordination schemes.

Keywords

Average effective marginal tax; Beggar-thy-neighbour; Capital controls; Cooperation; Cross-border shopping; Excise taxes; Flat tax; Marginal effective tax rate; Tax competition; Tax havens; Taxation of corporate profits; Transfer pricing

JEL Classifications

H2

Tax competition refers to strategic tax-setting in a non-cooperative game between jurisdictions – which can be countries, or states or provinces within a federation – with each setting some parameters of its tax system in relation to the taxes set by others.

Broadly read, this definition of tax competition encompasses essentially all tax policy, since every decision potentially requires some view as to the tax strategies formed in other jurisdictions. Typically though, the term is intended to focus on explicit interactions in tax-setting. And the central policy concern to which such interactions point is the potential for efficiency losses (or, as will be seen, gains) – and hence potential gains (or losses) from cooperation – to the extent that each policymaker ignores (or at least attaches less importance to) the impact of its own tax decisions on other jurisdictions, creating fiscal externalities between them.

These cross-border effects might, in principle, lead to taxes ending up too *high* from a collective perspective, rather than too low, as a result of 'tax exporting'. Most obviously, countries with power in the world market for some commodity may find it in their interest to explicitly tax those exports. Similar effects may also be at work in relation to corporate taxation: for example, to the extent that profits arising domestically accrue to foreigners, there is an incentive to tax them heavily (since the well-being of foreigners is presumably less valued than the welfare of domestic citizens).

But the main concern in this area has been with the potential for a 'race to the bottom' arising from the possibility that a cut in one country's tax rate will make other countries worse off – losing tax base and/or real economic activity. If each policymaker ignores these harmful cross-border effects, there is a risk of a 'beggar thy neighbour' situation in which tax rates end up generally too low in terms of the collective interest.

There are, though, two reasons why this may not pose as great a policy problem as it sounds. First, if policymakers are to some degree self-serving rather than wholly benevolent, then the constraints that tax competition imposes on them may on balance be beneficial for the citizenry. Second, even though governments do not cooperate explicitly, since the tax-setting game between them is played many times, they may find ways to tacitly sustain the cooperative outcome and so avoid inefficiencies.

Tax competition can affect many aspects of tax design, but the policy concerns it raises are naturally greatest where tax bases are most mobile. At an international level, this has meant a focus on the taxation of capital income and excises on such readily transportable and conventionally heavily taxed commodities as cigarettes and alcohol. For brevity, this article follows most of the literature (Wilson 1999) in focusing on these two, and moreover focuses, within the former, on corporate taxation.

Excise Tax Competition

The international norm is that commodity taxes are levied on the 'destination principle', meaning that tax is charged according to the jurisdiction in which consumption occurs: French wine consumed in the United Kingdom, for example, is taxed at the United Kingdom rate. This leaves some room for strategic tax setting (a point

developed by Friedlander and Vandendorpe 1968): prohibited from levying an import tariff on a fellow member of the European Union, the United Kingdom, for example, might be tempted to set a relatively high excise on wine, domestic and imported, so as to dampen import demand and generate a terms of trade benefit or as a means of rent shifting. The scope for tax competition is greatly increased, however, where – as a consequence of legal shopping across borders by consumers and smuggling (for brevity, we refer to both as 'cross-border shopping') – the destination principle proves difficult to enforce, so that commodities are in effect taxed according to where they are produced: the 'origin principle'.

With taxation on an origin basis – at least de facto, and to a potentially significant extent – countries have an incentive to set a relatively low tax rate in order to protect their tax base and perhaps gain base at the expense of others. And there are clearly instances in which cross-border shopping has been a significant influence in tax-setting. The classic example is the reduction in Canadian cigarette taxes in the early 1990s in response to widespread smuggling across the US border. Also, in the United Kingdom cross-border shopping with France has for many years been cited explicitly as constraining the rate applied to alcoholic drink. There is, it should be noted, a further complication with the excises. This is the risk that cigarettes will not even bear the tax charged by a low-tax country, but will simply be diverted to consumption without payment of any tax (Keen 2002a).

Corporate Tax Competition

Competition in capital income taxation is potentially most powerful when the tax is levied on a 'source' or 'territorial' basis (that is, on the income derived in particular jurisdictions) rather than on a 'residence' basis (on the income that residents in country derive from their income around the world). This is because under a pure residence system the only way that individuals or companies can take advantage of lower tax rates offered in other countries is by changing their residence: simply by relocating their investments to low-tax jurisdictions, or transfer pricing profits into subsidiaries located there, they would not ultimately avoid the taxes in their country of residence.

In practice, capital income taxes often have a significant element of source taxation even when formally levied on a residence basis. Individuals may simply locate their savings in low-tax jurisdictions and fail to report the proceeds, being especially secure in this when the source country provides a measure of banking secrecy. At corporate level, exemption may be explicit, with outright exemption of corporation's earnings from abroad: this is the case, for example, in the Netherlands and (for countries with which it has a double tax treaty) Canada. Or it may be implicit: while many countries, including the United States and United Kingdom, in principle apply the residence principle, their taxes typically apply only when a multinational's subsidiary abroad pays dividends to the parent, so that those taxes can be deferred (and hence reduced in present value) by delaying repatriation. This brings the system closer to one of de facto source taxation.

Countries that apply the residence principle find it is coming under increasing pressure, particularly for individuals but also for corporations. One sign of this has been the emergence of corporate inversions in the United States and elsewhere, with companies shifting their residence to low-tax jurisdictions. Another is the increased emphasis on controlled foreign corporation (CFC) rules, under which profits of subsidiaries earned abroad, typically in low-tax jurisdictions, may be brought into tax even if not repatriated to the parent.

Importantly, the impact of the corporate tax on business decisions depends on more than just the rate of the tax: it depends on what allowances are available for depreciation, financial costs and other expenses. The question then arises as to precisely which aspects of the corporate tax countries might compete over. There are three candidate tax rates.

First, attention has naturally focused on the headline statutory rate of corporation tax. This is especially relevant to firms' decisions regarding

income shifting, meaning essentially the use of devices other than real investment – transfer pricing, financial arrangements and so on – to shift taxable receipts from countries in which the statutory rate is high to those in which it is low, and deductible expenses in the opposite direction. Statutory rates have fallen substantially since the 1980s. In the Organisation for Economic Co-operation and Development (OECD) countries, the top rate of corporation tax fell from 41 per cent in 1986 to around 27 per cent in 2007. And this reduction in statutory rates has not been confined to developed countries: in sub-Saharan Africa, for example, it fell, on average, by about ten points over the 1990s (Keen and Simone 2004).

Second, decisions as to the level of investment in a given country depend on the marginal effective tax rate (METR), which is a summary measure of the combined impact of the statutory rate and the allowances available on the return that a firm must earn in order to provide investors with the after-tax return they require. This is an indicator of the effect of the tax system on the incentive to invest at the margin. If the METR is zero, for example, then the tax system has no impact on the marginal decision to invest even though it may collect revenue by taxing the return inframarginal investments. Strikingly, for the OECD countries the METR has remained broadly stable over the 1990s (Devereux et al. 2002) – which indicates that the reduction in statutory rates has been offset to some degree by a broadening of the corporate tax base.

Third, the choice as to the country in which to locate a given discrete investment (a factory, for example) depends on a comparison of the average effective marginal tax rate (AETR) in each, this quantity reflecting the present value of taxes to be paid – including on infra-marginal profits – over the life of the project. In practice, the AETR often tracks the statutory tax rate quite closely (Devereux and Griffith 2003), so that it too has fallen substantially over the 1980s.

The overall picture thus suggests that the developed countries have been competing aggressively over the statutory rate of corporation tax, consistent with a desire to benefit from (or prevent) income shifting, but have broadened the tax base sufficiently to leave overall incentives to real investment broadly unchanged.

In What Sense Is This Tax Competition?

It could be that these trends simply reflect common developments in a range of countries rather than direct interaction between them – perhaps a common desire to improve incentives by reducing top rates of personal income tax, with the reduction in the corporate tax an adjunct to this, driven by the need to prevent disincorporation. Or – consistent with the basic notion of tax competition above – it could be a form of 'yardstick competition', with governments perceiving that their electorates assess them in part by comparing them with neighbouring countries, and so use the tax system to signal a supportive attitude to business or wider competence in economic management. Distinguishing empirically between these two hypotheses – correlation due to competition for tax base or real investment, versus correlation due to common shocks or yardstick competition – is not easy. Brueckner (2003) provides a review of the empirical literature in this area. In the international context, the emerging empirical evidence does seem to suggest that tax competition in pursuit of tax base or investment is important: Devereux et al. (2003), for example, find that the correlation between corporate tax rates became greater as capital controls became weaker, which would not be the case if the correlation were due to common domestic shocks. More anecdotally, one recent sign of tax competition closer to yardstick form than to competition for mobile base may be the spread, in the last few years, of 'flat tax' systems characterized by low tax rates on personal income: while one can see an argument for a low corporate tax rate to attract mobile capital, it is hard to see a similar base gain from setting the tax rate on labour income at the same, low rate.

This account of developments in the various corporate tax rates suggests that the pattern of any corporate tax competition has been quite complex. Why is it that countries have apparently competed

over the tax rate but not over the base? One possibility is that this enables them to attract the more mobile and profitable investments without overly jeopardizing revenue from the less mobile corporate tax base. But there is no direct evidence for this, so the pattern of competition remains something of a puzzle.

Revenue Effects

There has been much talk of tax competition leading to the erosion – perhaps the elimination – of revenue from the corporate tax, and from capital income taxes more generally.

For the developed countries as a group, however, this has not happened. Indeed, for them corporate tax revenues since the 1990s have if anything increased. In the EU15, for example, corporate tax revenues rose from 2.5 per cent of GDP in 1990 to 3.2 per cent in 2004. Country experiences vary, however. In Japan, corporate tax revenue has decreased over the same period from 6.5 to 3.8 per cent of GDP, and in Germany from 2.3 to 0.6 per cent. While these revenue figures need to be interpreted with great caution, there is no sign (as of 2007) of a collapse.

Quite why revenue has held up in the developed countries is not fully understood. Base-broadening measures have in many cases been adopted alongside the reduction in statutory rates, as seen above, but it is not clear that these have been sufficient to provide a full explanation. For the United Kingdom, there is evidence (Devereux and Klemm 2005) that while base-broadening played some role, the buoyancy of corporate tax revenues has reflected also growth of the corporate sector, especially the financial sector, that did not seem to be largely attributable to the tax itself. It may well be that in other countries too – though not, it seems, all – an increase in the share of profits in GDP is part of the reason. For the United States, Auerbach (2006) argues that the resilience of corporate tax revenue reflects increased volatility of profits: the asymmetric treatment of positive and negative profits (the former taxed, the latter attracting no rebate but only carry forward) implies that this would lead to an increase in expected tax revenues. It is by no means clear, of course, that any such trends – and, hence, the resilience of corporate tax revenues – can be expected to continue.

There is tentative but emerging evidence of quite a different story in developing countries. These, too, have seen a marked reduction in statutory rates of tax. But this has not been offset by an expansion of the base – revenues in relation to GDP appear to have been falling, by about one-fifth in the poorest of them since the 1990s. Given their pressing needs for revenue and greater reliance on corporate taxes – around 13 per cent of total revenues at the start of the 1990s, compared to about nine per cent for the high-income countries – it may be that corporate tax competition is a more pressing concern for developing countries than for developed.

Basic Welfare Concerns

The policy concern raised by tax competition is that the failure to coordinate – each jurisdiction attaching relatively little importance to the impact of its decisions on others – will lead to a worse outcome than could be obtained from some form of cooperation in tax-setting. The fear is that the revenue pressures which emerge will lead either to reduced government expenditure or reliance on other tax instruments, notably taxes and social contributions on labour income, that are more distortionary and/or less coherent with equity objectives. In the limit (and leaving aside location- specific rents, most classically from natural resources), the corporate tax, for instance, would be reduced to a benefit tax – that is, one which simply charges companies for the value of the services they receive. Competition for mobile capital may also distort the composition of public spending, as well as its level, with too much focus on public infrastructure complementary with capital and too little on public expenditure on items of benefit to consumers: too many airports, not enough libraries. The argument is developed by Keen and Marchand (1997), and there is emerging evidence to suggest it is of some real importance.

Winners and Losers

This lack of coordination does not mean, however, that *all* countries will be worse off as a result of tax competition. The literature strongly suggests, in particular, that small countries are likely to set particularly low tax rates, and in doing so to be better off than they would be under schemes of cooperation that do not explicitly involve the large countries paying them transfers (Kanbur and Keen 1993). This is because in setting low tax rates they have little to lose in terms of revenue from their domestic tax base but a lot to gain (both in revenue terms and, perhaps, for the scale and profitability of their financial sector) from attracting tax base initially located abroad. And of course many tax havens are indeed small jurisdictions.

Thus coordination is not necessarily Pareto-improving: inducing all countries to participate may require transfers to those who win from tax competition and/or the exercise or threat of some form of punishment for non-cooperation.

Tax Competition may be Good: 'Taming the Beast'

Some see tax competition as a good thing, providing a market discipline that serves to limit the size of government, supplementing inadequate constitutional and electoral constraints. In this view, developed first by Buchanan and others taking the view of government as 'Leviathan', the inefficiencies from non-coordination reduce the welfare of the policymakers themselves but increase that of the citizenry. There is another more subtle argument to the effect that tax competition may be beneficial even when policymakers are wholly benevolent. This is because the possibility of capital flight to low-tax countries provides a way in which others can commit not to impose heavy *ex post* taxes on savings once they have been made (so overcoming a basic time consistency problem in taxing capital income): see Kehoe (1989).

Much of the policy debate on the desirability or otherwise of tax competition quickly becomes a sterile statement of ideology, with it being seen as either wholly bad or wholly good. But there is a strand of the literature that has sought to better understand the trade-off at issue, between improved efficiency of the tax system as a result of coordination and the possibility that some or all of the additional revenue this can be used to generate will be wasted. One model leads to a simple test: a coordinated increase in tax rates from the non-cooperative equilibrium increases the citizens' welfare if and only if $\lambda < MDL/(1 + MDL)$, where λ is the proportion of public expenditure that is wasted and MDL is the marginal deadweight loss from raising an additional euro of revenue (Edwards and Keen 1996). This at least enables one to narrow down the scope of policy disagreement. Suppose, for example, it is agreed that the marginal deadweight loss from taxation is at least 15 cents per euro. Then all those who believe that government wastes no more than 13 cents per euro of its spending, at the margin, should agree that a small coordinated increase in the corporate tax rate would be beneficial. The key difficulty remains, however, of determining what proportion of spending is 'wasted' – or indeed what 'waste' means in this context, since different people may clearly take different views, for example, of the social value of spending that is essentially redistributional.

Besley and Smart (2007) provide a more subtle analysis of these political economy issues in a model of electoral competition, in which candidates may be either Leviathans or benevolent. The results that emerge are more nuanced, but include, importantly, the observation that intensified tax competition is likely to be beneficial to the citizenry only if pure Leviathans are sufficiently rare: otherwise the electoral process offers them little respite from exploitation.

Quantification

There have been a few attempts to estimate the welfare costs of tax competition, using computable general equilibrium models. On the corporate tax side, Parry (2003) finds the loss to be relatively small: about three per cent of capital tax revenues, and even less when a modest degree of

'Leviathanism' is present. Sørensen (2004a) also finds relatively small gains: less than one per cent of GDP. On the excises, Keen (2002a) stresses the difficulty of inferring the extent of any problem from the extent of observed cross-border shopping: tax competition could be so fierce, for example, that this is zero in equilibrium and yet a welfare loss is suffered from the inefficiently low equilibrium tax rate. But illustrative calculations in this case also suggest a relatively modest welfare loss: rarely more than two per cent of tax revenue.

Ensuring that All Participating Countries Gain

The first problem in coordinating taxes is to ensure that all participants gain. For the EU, this is explicit in the unanimity rules; in other contexts, it is implicit in the national sovereignty of potential participants.

In principle, the inefficiency being addressed implies that all could gain from coordination *if* accompanied by compensating transfers to some participants. This means, more specifically, making payments to any countries that gain from tax competition. This appears unlikely in practice, because of the appearance of rewarding those who are gaining at the expense of others. It may be that the best way to deal with this is by combining coordination with other measures, perhaps not in the tax area, that tend to benefit the winners: this is the 'Hicksian optimism' that by adopting a series of reforms that are potentially Pareto-improving one will arrive at an actual Pareto improvement. The alternative is the exercise of threat.

The 'Third Country' Problem

Coordination among a subset of countries may be undermined if other countries do not also participate, a point that has been explicit, for example, in the negotiations leading to the EU Savings Directive. This raises the same issues of compensation and cajoling just discussed. The third country problem does not mean that coordination among that subset alone would actually leave them worse off (Konrad and Schjelderup 1999). Simulation exercises do though suggest that, when capital mobility is high, the welfare gains may be far smaller when only a subset of countries participate (Sørensen 2004a).

Full Harmonization

There are circumstances in which harmonization – a term that has come to mean complete uniformity not only of rates but also of bases (especially problematic for the corporate tax, but not a trivial issue for the excises either) – is collectively beneficial. Keen (1989), for example, shows that starting from a Nash equilibrium in the setting of destination-based commodity taxes, convergence to an appropriately weighted average of the initial tax rates is Pareto-improving. This result supposes, however, that taxes are used only for strategic reasons: by adding a revenue motive the conditions for Pareto gains become much stronger, as shown in Lockwood (1997). In practice, moreover, such full convergence is not only politically unlikely but is overly restrictive as a means of dealing with coordination failure. The search has been for looser measures of coordination.

A Minimum Tax Rate

Minimum rates are in principle an attractive way of limiting any 'race to the bottom', leaving potentially considerable leeway for national discretion in tax-setting. Even those countries initially below the minimum, and so required to raise their tax rates, may benefit from the adoption of a minimum tax: this is because when they raise their rates, countries above the minimum will be less threatened by their low rates and so will tend to set higher rates than they otherwise would. This increase in the rates set by those above the minimum reduces the damage to those forced to raise their rates, and may even cause the latter to gain (Kanbur and Keen 1993). In this way, imposing a minimum rate may be Pareto-improving.

All this assumes, however, that countries compete only over the statutory rate of tax. That may be a reasonable assumption in relation to excise taxation (for which the EU has indeed adopted minimum rates), since there is relatively little scope for game-playing on the base itself. For the corporate tax, however, the danger is that unless there is also agreement on a common base for the corporate tax, countries may instead compete by narrowing their tax bases, offering more generous allowances for investment, and so on. Tax competition would thus simply manifest itself in a different form.

Formula Apportionment

This refers to a corporate tax system – of the kind operated by the states in the United States and the provinces of Canada – under which the profits earned by multi-jurisdictional enterprises are allocated across those jurisdictions by means of some formula intended to capture the extent of its activities in each, and each jurisdiction then taxes the profits allocated to it at whatever rate it chooses.

The advantage of such a scheme is that it eliminates the incentive for multinationals to move profits between jurisdictions by transfer-pricing or financial arrangements, since these have no effect on aggregate profits and hence also no effect on the taxes charged in each jurisdiction. This in turn means that the jurisdictions have no incentive to set low tax rates in order to encourage such income-shifting.

Companies will have an incentive to distort their activities across jurisdictions, however, in so far as this affects the weights by which their profits are allocated across jurisdictions. This makes it important, for example, that (in contrast to common practice in the United States) the capital stock should *not* enter the formula. If it does, companies will have an incentive to invest in jurisdictions with a low tax rate; and there will be an incentive to offer low tax rates in order to attract such investment. Indeed, the net effect could be that tax competition is actually worse under formula apportionment than under separate accounting (Sørensen 2004b): encouraging a firm to invest a little more in a country produces revenue proportional to the *marginal* return on that investment under separate accounting, but proportional to the firm's *average* profit under formula apportionment. Since average returns tend to exceed marginal, this makes it more tempting to attract capital by offering a low tax rate. The general difficulty here is that under formula apportionment the corporate tax becomes to some degree a tax on whatever is used to define the weights. Thus, similar but perhaps less problematic effects arise with using sales and some measure of employment in the weights, these being the other main candidates.

Ring-fenced Corporate Tax Regimes: Good or Bad?

A recurrent theme in attempts to identify especially 'harmful' aspects of corporate tax competition is the idea that special schemes which are 'ring-fenced' in the sense of being restricted to particular investors or sectors are especially damaging.

This may not be correct (Keen 2002b; Janeba and Smart 2003). Allowing countries to compete very aggressively over particularly mobile aspects of the corporate tax base while maintaining higher rates on the rest may lead to an outcome that is better for all concerned than one in which they are required to set the same tax rate on all parts of the base. The reason is simple: it may ultimately be less damaging for countries to compete very aggressively over a narrow base than to compete less aggressively over a wider one.

Codes of Conduct

The attempt to prevent the spread of, and to roll back, particular practices by means of non-binding rules of the game has been a remarkable development in international taxation over recent years, with such a code of conduct being adopted in the EU and a similar approach being adopted as part of the OECD's harmful tax practice project. The question arises as to how much further such an approach

can be pushed. For by focusing on special schemes and excluding general levels of corporate taxation – and even putting aside the possible reservation on ring-fencing just mentioned – the codes arguably miss the central issue: too low a general level of corporate taxation.

Information Sharing: Reinforcing Destination and Residence Principles

A quite different strategy is to limit the scope for tax competition by seeking to strengthen the application of the destination and residence principles (for commodity and income taxes respectively). The former is difficult to do given the general trend towards seeking fewer border formalities rather than more, an effect amplified by the increased importance of international services that cannot be taxed as they cross the border. For the income tax, strengthening resident taxation would involve, in particular, limiting the scope for deferral of taxes by leaving earnings aboard, in turn entailing an extension and firmer enforcement of CFC rules.

Another key measure that has been the focus of much attention in recent years is the strengthening of international information exchange on tax matters. In relation to capital income, this is a key element of the EU savings directive (which requires member states to either provide information to others on the interest income earned by their residents or levy a withholding tax, most of the proceeds being transferred to the country of residence) and the OECD's harmful-tax project.

It may seem obvious that low-tax countries can only lose from sharing information, through a reduction in their tax base and associated activities. Bacchetta and Espinosa (1995), however, show that they may benefit from a strategic effect of such sharing: for the higher-tax countries will then be inclined to set higher rates than they otherwise would (prospective outflows being diminished by the prospect of discovery abroad), which also enables the low-tax country to raise its tax rate too. Building on this insight, Huizinga and Nielsen (2003) explore the choice between information sharing and the use of withholding taxes in a repeated game, while Keen and Ligthart (2006) show that, if the difference in country size (and hence noncooperative tax rates) is large enough, then sharing some of the revenue raised as a result of shared information with the source country (contrary to standard practice) has an adverse impact on total revenue raised but provides a device for securing a Pareto gain. How powerful the strategic effects of information exchange are likely to be, and indeed how effective such measures are likely to be in a technical sense (given, not least, the absence of common taxpayer identification numbers across countries) remains to be seen.

Tax Competition in Federations

The discussion so far has been concerned with 'horizontal' tax competition, in the sense that the interaction has been between jurisdictions with their own distinct tax bases. In federal systems, however, different levels of government commonly share tax bases: this may be explicit – in the United States, for example, corporate income is taxed at both federal and state level – or implicit (the base of a federal income tax, for instance, would overlap substantially with that of a state sales tax). The latter gives rise to 'vertical' tax competition, which in itself might be expected to lead to taxes that are too high from the collective perspective: a lower-level government that increases its own tax rate is liable to take less than full account of the impact on federal revenues of the consequent contraction of the shared tax base.

Two aspects of vertical tax externalities have received particular attention. First, how does tax-setting at the two levels interact? Theory leaves this indeterminate: intuitively, the optimal tax rate set by a lower-level jurisdiction is likely to depend (inversely) on the elasticity of the base, the effect of a change in the federal tax rate is in principle typically unclear (see, for example, Keen 1998). Empirically, Hayashi and Boadway (2001), for example, find that higher federal rates of corporate taxation in Canada are associated with lower provincial tax rates. Second, how

does the interplay between horizontal and vertical externalities play out (with the former pointing in most models to tax rates being too low, and the latter to their being too high)? In a model of capital income tax competition, Keen and Kotsogiannis (2002) show that this turns on the relative magnitudes of the elasticities of the demand for capital (which shapes the aggressiveness of horizontal tax competition) and of the supply of savings (shaping the responsiveness of the shared tax base). In a model of excise taxation, capturing both directions of externality, Devereux et al. (2007) show that the balance between the two depends on the ease of cross-border arbitrage and the price elasticity of demand. Using data for the US states, they also find evidence of significant vertical interactions in the setting of gasoline taxes, with the federal tax tending to be positively associated with state taxes.

Conclusion

The marked reduction in statutory corporate tax rates over the 1980s, which seems to be largely if perhaps not entirely attributable to international tax competition, has not been matched by a similar reduction in corporate tax revenues – at least in developed countries. Quite why remains something of a puzzle, and the possibility of more marked reductions in the future cannot be ruled out. The welfare significance of these developments also remains a matter of dispute, most fundamentally because the view that tax competition may provide a useful discipline on government has not been developed to a point at which it has firm empirical substance. The case for coordination thus remains uneasy, and is perhaps strongest in developing countries given both the more apparent impact on revenues and the clearer need there for stronger revenue mobilization. If coordination is sought, the key difficulty is to ensure that it takes a form from which all participants gain – in the absence of explicit transfers, this is not easy to assure, and may require packaging measures within some broader agreement. Recent policy initiatives have focused more on administrative measures than on substantive policy restrictions: it remains unclear how effectively these will deal with the underlying problems that motivate their use.

See Also

- ▶ Excise Taxes
- ▶ Tax Havens
- ▶ Taxation of Corporate Profits
- ▶ Taxation of Foreign Income

Bibliography

Auerbach, A. 2006. The future of capital income taxation. *Fiscal Studies* 27: 399–420.

Bacchetta, P., and M.P. Espinosa. 1995. Information sharing and tax competition among governments. *Journal of International Economics* 39: 103–121.

Besley, T., and M. Smart. 2007. Fiscal restraints and voter welfare. *Journal of Public Economics* 91: 755–773.

Brueckner, J. 2003. Strategic interaction among governments: An overview of empirical studies. *International Regional Science Review* 26: 175–188.

Devereux, M., and A. Klemm. 2005. Why has the U.K. corporation tax raised so much revenue? *Fiscal Studies* 25: 367–388.

Devereux, M., and R. Griffith. 2003. Evaluating tax policy for location decisions. *International Tax and Public Finance* 10: 107–126.

Devereux, M., B. Lockwood, and M. Redoano. 2007. Horizontal and vertical indirect tax competition: Theory and evidence from the USA. *Journal of Public Economics* 91: 451–479.

Devereux, M., R. Griffith, and A. Klemm. 2002. Corporate income tax reforms and tax competition. *Economic Policy* 35: 451–495.

Devereux, M., B. Lockwood, and M. Redoano. 2003. Capital account liberalization and corporate taxes. IMF Working Paper 03/180. Online. Available at http://www.imf.org/external/pubs/ft/wp/2003/wp03180.pdf. Accessed 3 July 2007.

Edwards, J.S.S., and M.J. Keen. 1996. Tax competition and leviathan. *European Economic Review* 40: 113–143.

Friedlander, A.F., and A.L. Vandendorpe. 1968. Excise taxes and the gains from trade. *Journal of Political Economy* 76: 1058–1068.

Hayashi, M., and R. Boadway. 2001. An empirical analysis of intergovernmental tax interaction. *Canadian Journal of Economics* 34: 481–503.

Huizinga, H., and S.B. Nielsen. 2003. Withholding taxes or information exchange: The taxation of international interest flows. *Journal of Public Economics* 87: 39–72.

Janeba, E., and M. Smart. 2003. Is targeted tax competition less harmful than its remedies? *International Tax and Public Finance* 10: 259–280.
Kanbur, R., and M. Keen. 1993. Jeux Sans Frontières: Tax competition and tax coordination when countries differ in size. *American Economic Review* 83: 877–892.
Keen, M. 1989. Pareto-improving indirect tax harmonization. *European Economic Review* 33: 1–12.
Keen, M. 1998. Vertical tax externalities in the theory of fiscal federalism. IMF Staff Papers 45, 454–85. Online. Available at http://www.imf.org/external/Pubs/FT/staffp/1998/09-98/pdf/keen.pdf. Accessed 3 July 2007.
Keen, M. 2002a. Some international issues in commodity taxation. *Swedish Economic Policy Review* 9: 11–45.
Keen, M. 2002b. Preferential regimes can make tax competition less harmful. *National Tax Journal* 54: 757–762.
Keen, M., and C. Kotsogiannis. 2002. Does federalism lead to excessively high taxes? *American Economic Review* 92: 363–370.
Keen, M., and J. Ligthart. 2006. Incentives and information exchange in international taxation. *International Tax and Public Finance* 13: 163–180.
Keen, M., and M. Marchand. 1997. Fiscal competition and the pattern of public spending. *Journal of Public Economics* 66: 33–53.
Keen, M., and A. Simone. 2004. Tax policy in developing countries: some lessons from the 1990s, and some challenges ahead. In *Helping countries develop: The role of the fiscal policy*, ed. S. Gupta, B. Clements, and G. Inchauste. Washington, DC: IMF.
Kehoe, P. 1989. Policy cooperation among benevolent governments may be undesirable. *Review of Economic Studies* 56: 289–296.
Konrad, K., and G. Schjelderup. 1999. Fortress building in global tax competition. *Journal of Urban Economics* 46: 156–167.
Lockwood, B. 1997. Can international commodity tax harmonization be Pareto-improving when governments supply public goods? *Journal of International Economics* 43: 387–408.
Mintz, J., and H. Tulkens. 1986. Commodity tax competition between member states of a federation. *Journal of Public Economics* 29: 133–172.
Parry, I.W.H. 2003. How large are the welfare costs of tax competition? *Journal of Urban Economics* 54: 39–60.
Sørensen, P.B. 2004a. International tax coordination: Regionalism versus globalism. *Journal of Public Economics* 88: 1187–1214.
Sørensen, P.B. 2004b. Company tax reform in the European Union. *International Tax and Public Finance* 11: 91–115.
Wilson, J.D. 1986. A theory of interregional tax competition. *Journal of Urban Economics* 19: 296–315.
Wilson, J.D. 1999. Theories of tax competition. *National Tax Journal* 52: 269–304.
Zodrow, G., and P. Mieszkowski. 1986. Pigou, property taxation and the underprovision of public goods. *Journal of Urban Economics* 19: 356–370.

Tax Compliance and Tax Evasion

Joel Slemrod

Abstract

Tax evasion is widespread, always has been, and probably always will be. Variations in duty and honesty can explain some of the across-individual and, perhaps, across-country heterogeneity of evasion. But the stark differences in compliance rates across taxable items that line up closely with detection rates suggest strongly that deterrence is a power factor in evasion decisions. Although the normative theory of taxation has been extended to tax system instruments such as the intensity of enforcement, the empirical knowledge for operationalizing these rules is sparse.

Keywords

Civic virtue; Deterrence; Reciprocity; Tax compliance; Tax evasion; Tax incidence

JEL Classifications

H2

No government can announce a tax system and then rely on taxpayers' sense of duty to remit what is owed. Some dutiful people will undoubtedly pay what they owe, but many others will not. Over time the ranks of the dutiful will shrink, as they see how they are being taken advantage of by the others. Thus, paying taxes must be made a legal responsibility of citizens, with penalties attendant on noncompliance. But even in the face of those penalties, substantial tax evasion exists – and always has.

Determining the extent of evasion is not straightforward, for obvious reasons. Because tax evasion is both personally sensitive and potentially incriminating, selfreports are vulnerable to substantial underreporting. Moreover, the dividing line between illegal tax evasion and legal tax avoidance is blurry. Under US law, tax evasion

refers to a case in which a person, through commission of fraud, unlawfully pays less tax than the law mandates. Tax evasion is a criminal offence under federal and state statutes, subjecting a person convicted to a prison sentence, a fine, or both. An overt act is necessary to give rise to the crime of income tax evasion; therefore, the government must show wilfulness and an affirmative act intended to mislead. Some tax understatement is, however, inadvertent error, due to ignorance of or confusion about the tax law (as is some overpayment of taxes). Although the theoretical models of this issue generally refer to wilful understatement of tax liability, the empirical analyses cannot precisely identify the taxpayers' intent and therefore cannot precisely separate the wilful from the inadvertent. Nor can they, in complicated areas of the tax law, precisely distinguish the illegal from the legal.

The most careful and comprehensive estimates of the extent and nature of tax noncompliance anywhere in the world have been made for the federal taxes that the US Internal Revenue Service (IRS). The IRS comes up with its estimates by combining information from random intensive audits with information obtained from ongoing enforcement activities and special studies about sources of income, such as tips and cash earnings of informal suppliers like nannies and housepainters, that can be difficult to uncover even in an intensive audit.

The latest tax gap estimate, released in February 2006 (IRS 2006) but pertaining to the 2001 tax year, estimated the overall gross tax gap estimate to be 345 billion dollars, which amounts to 16.3 per cent of estimated actual (paid plus unpaid) tax liability. Of the 345 billion dollar estimate, the IRS expects to recover 55 billion dollars, resulting in a 'net tax gap' – that is, the tax not collected – for tax year 2001 of 290 billion dollars, which is 13.7 per cent of the tax that should have been reported.

About two-thirds of all underreporting happens on the individual income tax; the corporation income tax makes up slightly more than ten per cent and the payroll tax gap makes up about one-fifth of total underreporting. For the individual income tax, understated income, as opposed to overstating of exemptions, deductions, adjustments, and credits, accounts for over 80 per cent of underreporting of tax. Business income, rather than wages or investment income, accounts for about two-thirds of the understated individual income. Taxpayers who were required to file an individual tax return, but did not, accounted for slightly less than ten per cent of the gap.

There are wide variations in the rate of misreporting as a percentage of actual income by type of income (or offset). Only one per cent of wages and salaries and four per cent of taxable interest and dividends are underreported. In large part this is because wages and salaries, as well as interest and dividends, must be reported to the IRS by those who pay them; in addition, wages and salaries are subject to employer withholding. Self-employment business income is not subject to information reports or withholding, and its estimated noncompliance rate is sharply higher. An estimated 57 per cent of non-farm proprietor income is not reported – 68 billion dollars – which by itself accounts for more than a third of the total estimated underreporting for the individual income tax. All in all, over half of underreporting is attributable to the underreporting of business income, of which non-farm proprietor income is the largest component.

All in all, there is substantial evidence that the extent of evasion for sole proprietor income is high compared to such income sources as wages, salaries, interest and dividends, and may be more than half of true income. Other components of taxable income for which information reports are nonexistent or of limited value, such as other non-wage income and tax credits, also have relatively high estimated misreporting rates. The IRS reports (IRS 2006) that the net misreporting rate is 53.9, 8.5, and 4.5 per cent for income types subject to 'little or no,' 'some,' and 'substantial' information reporting, respectively, and is just 1.2 per cent for those amounts subject to both withholding and substantial information reporting.

Little is known about how the level of noncompliance, and its proportion to actual income, varies by income class. One study based on IRS audit data for 1988 suggested that higher-income people evade *less*, in relation to the size of their

true income, than those with lower incomes, but for a number of reasons this study is not conclusive. Other studies suggest that married filers, taxpayers younger than 65, and men have significantly higher average levels of noncompliance than others. Within any group defined by income, age, or other demographic category, there are some who evade, some who do not, and even some who overstate tax liability. It is not known to what extent this heterogeneity is explained by different 'tastes' for evasion or different opportunities to evade.

Noncompliance is also a factor with businesses, both in their role as withholding agents for taxes that are not statutorily levied on businesses, and also for taxes that are levied on businesses, such as the corporation income tax. Based largely on operational data, the IRS estimates that noncompliance with the corporation income tax in 2001 was 30 billion dollars, which corresponds to a noncompliance rate of 17 percent. Of this 30 billion dollars, noncompliance by corporations with over 10 million dollars in assets make up 25 billion. But the estimated noncompliance rate of the larger companies is lower, 14 per cent compared to 29 per cent for corporations with less than 10 million dollars of assets. Because these estimates are largely based on deficiencies proposed by the examination teams of operational audits, and because most big corporations are routinely audited, these tax gap estimates are subject to several caveats. Because of the complexity of the tax law, exactly what is actual tax liability – and therefore what is actual tax noncompliance – is often not clear. In any given audit, some noncompliance may be missed, and there will also be mistakes in characterizing as noncompliance what is legitimate tax planning. Knowing that the resolution of the ultimate tax liability is often a long process of negotiation, the tax liability according to the originally filed return, as well as the initial deficiency assessed by the examination team, may be partly a tactical 'opening bid' that is neither party's best estimate of the 'true' tax liability.

It is difficult to compare the magnitude and nature of tax evasion in the United States with other countries, in part because no other country has undertaken a broadbased analysis of tax evasion like that undertaken in the United States. Based on less extensive analysis, the Swedish Tax Agency has estimated the total gap as a percentage of taxes at eight per cent in 2000. Although no official estimate for the United Kingdom has been released, a government document has speculated that it is likely that the United Kingdom has a tax gap of a similar magnitude to that of Sweden and the United States. Many studies suggest that noncompliance rates in developing countries are considerably higher.

Economics models have tried to put these facts into a coherent model. The standard economics framework for considering an individual's choice of whether and how much to evade taxes is a deterrence model in which taxpayers make these decisions in the same way they would approach any risky decision or gamble – by maximizing expected utility – and are influenced by possible penalties no differently than any other contingent cost. Optimal tax evasion depends on the chance of getting caught and penalized, the size of the penalty for evasion, and the individual's degree of risk aversion.

Attempts to empirically verify the predictions of the deterrence model of tax evasion have focused on the effect on evasion of enforcement intensity and the level of tax rates, but have been plagued by the same measurement issues that arise in assessing the magnitude of tax noncompliance. Perhaps the most compelling empirical support for the deterrence model is the cross-sectional variation in noncompliance rates across types of income and deductions. Line item by line item, there is a clear negative correlation between the noncompliance rate and the presence of enforcement mechanisms such as information reporting and employer withholding. A striking example of the link from a lack of deterrence to tax compliance involves state use taxes, which are due on sales purchased from out-of-state vendors but consumed in the state of residence. These taxes are largely unenforceable (except perhaps for some expensive items like cars), and noncompliance rates are in the range of 90 per cent. The effect on noncompliance of the penalty for detected evasion, as distinct from the probability that a given act of noncompliance will be subject

to punishment, has not been compellingly established empirically.

Although the deterrence model has dominated the economics literature, some have argued that it predicts a compliance rate much lower than what we actually observe, and that factors such as duty and reciprocal altruism can explain this. Some have argued that many taxpayers comply with tax liabilities because of 'civic virtue', and that more punitive enforcement policies may crowd out such intrinsic motivation by making people feel that they pay taxes because they have to, rather than because they want to. Others argue that tax evasion decisions depend on perceptions of the fairness of the tax system or what the government uses tax revenues for. But such individual judgements can be complex; for example, expenditures on warfare might be tolerated in a patriotic period, but rejected during another period characterized by anti-militarism. These patterns suggest that a form of reciprocal altruism may be at work where taxpayer behaviour depends on the behaviour, motivations, and intentions not of any subset of particular individuals, but of the government itself. In support of this view, surveys show a positive relationship across countries between attitudes towards tax evasion and professed trust in government.

There is, however, no clear evidence that tax compliance behaviour can be easily manipulated by the government to lower the cost of raising resources. Appeals to patriotism to induce citizens to pay their taxes (and, often, buy war bonds) are common; the US Secretary of Treasury during the First World War, William Gibbs McAdoo, referred to these campaigns as 'capitalizing patriotism'. That such campaigns are successful during ordinary (non-war) times in convincing taxpayers to forego the cost-benefit calculus and comply has not been compellingly demonstrated. Recent randomized field experiments in the state of Minnesota and in Switzerland have found no evidence that appeals to taxpayers' consciences, stressing either the beneficial effects of tax-funded projects or conveying the message that most taxpayers were compliant, had a significant effect on compliance.

The difficulties of separating out whether people pay their taxes because they feel they 'ought to' or whether they fear the penalties attendant to not doing so is well illustrated by some evidence from a recent survey sponsored by the Internal Revenue Service (IRS Oversight Board 2006). While 96 per cent of those surveyed in 2005 mostly or completely agreed that 'It is every American's civic duty to pay their fair share of taxes', 62 per cent also said that 'fear of an audit' had a great deal or somewhat of an influence on whether they report and pay their taxes 'honestly'.

Tax evasion has policy implications because it affects the distribution of the tax burden as well as the resource cost of raising taxes. Variations in compliance rates by income class can to some extent be offset by adjustments in the rate schedule, but it is practically impossible to offset variations within an income class, so that evasion creates horizontal inequity because equally well-off people end up with different tax burdens.

Tax evasion also imposes efficiency costs. The most obvious are the resources taxpayers expend to implement and camouflage noncompliance, and the resources the tax authority expends to address this. In addition, when the tax system is otherwise close to optimal it provides a socially inefficient incentive to engage in those activities for which it is relatively easy to evade taxes. For example, because the income from house painting can be done on a cash basis and is therefore harder to detect, this occupation is more attractive than otherwise. Although a supply of eager and cheap house painters undoubtedly is greeted warmly by prospective buyers of that service, the work of the extra people drawn to house painting, or any activity that facilitates tax evasion, would have higher value in some alternative occupation.

The same argument applies to self-employment generally, as the enhanced opportunity for noncompliance inefficiently attracts people who would otherwise be employees. The opportunity for noncompliance can distort resource allocation in a variety of other ways, such as causing companies that otherwise would not find it attractive to set up a financial subsidiary, or set up operations in a tax haven, to facilitate or camouflage abusive avoidance or evasion.

The mere presence of tax evasion does not imply a failure of policy. Just as it is not optimal

to station a police officer at each street corner to eliminate robbery and jaywalking, it is not optimal to completely eliminate tax evasion. Recognizing tax evasion introduces a new set of policy instruments whose optimal setting is at issue, such as the extent of audit coverage, the strategy for choosing audit targets, and the penalty imposed on detected evasion. It also invites a rethinking of standard taxation problems.

One important issue is how many resources to devote to enforcing the tax laws. One superficially intuitive rule – increase the probability of detection until the marginal increase of revenue thus generated equals the marginal resource cost of so doing – is incorrect. Although the cost of hiring more auditors, buying better computers and the like is a true resource cost, the revenue brought in does not represent a net gain to the economy, but rather a transfer from private (noncompliant) citizens to the government. The correct rule equates the marginal social benefit of reduced evasion, which is not well measured by the increased revenue, to the marginal resource cost. The distinction suggests that unregulated privatization of tax enforcement, in which profit-maximizing firms would maximize revenue collection net of costs, would lead to socially inefficient overspending on enforcement. The social benefit is related to the reduced risk bearing that comes with reduced tax evasion and a reduction in the resource misallocations generated by evasion. Some have suggested that the basic framework of social welfare maximization is inappropriate, and have argued that there should be a specific social welfare discount applied to the utility of those who are found to be guilty of tax evasion and thus are known to be 'antisocial'; the standard normative model applies no such discount, so that noncompliant taxpayers do not per se receive a lower social welfare weight than compliant taxpayers.

No one has yet compellingly translated this theoretical characterization of optimal enforcement into a statement about how much evasion should be tolerated. But its implication for interpretation of the tax gap is clear and was stated by former IRS Commissioner Lawrence Gibbs, who said that the tax gap estimates are not intended to be measures of the potential for additional enforcement yields because some would not be 'cost-effective' to collect. An economist would substitute the term 'socially optimal' for 'cost-effective,' but the spirit of Gibbs's remark is essentially correct. Just as there is an important difference between oil reserves and 'economically recoverable' oil reserves, there is a difference between tax evasion and economically (read optimally) recoverable tax evasion.

The normative theory has not yet made much progress in guiding policy regarding the key tools of tax administration, especially the role of information reporting by arms-length parties. The ability of the IRS to rely on reports by firms about wages and salaries paid to employees explains why the (optimal) noncompliance rate of labour income is so much lower than for self-employment income, for which no such information reports exist. The ability to match firm-to-firm sales is touted by advocates as a major administrative advantage of value-added taxes, and the difficulty of monitoring firm-to-consumer sales and to distinguish them from firm-to-firm sales has been noted as the Achilles heel of administering a retail sales tax. Overall, when relatively disinterested third parties can be required to provide information, as they are with wages and salaries, high compliance rates can be achieved at fairly low cost. But when there are only interested parties involved, an alternative mechanism must be found – such as the requirement in an invoice-credit value added tax that taxes on input purchases can be deducted only if the seller produces an invoice for taxes remitted – or else compliance will be low in the absence of costly auditing.

The ubiquity and importance of evasion call into question one of the canons of undergraduate public finance textbooks – that the incidence and efficiency of taxes does not in the long run depend on which side of the market the tax is levied. Once the reality of tax evasion is recognized, the incidence and efficiency of a tax system may depend critically on which side of the market *remits* the tax to the government and which side must report its transactions to the government. A uniform value-added tax and a uniform national retail sales tax may look identical in a world of no

evasion or administrative costs, but have very different effects in the real world.

Tax evasion is widespread, always has been, and probably always will be. Variations in duty and honesty can explain some of the across-individual and, perhaps, across-country heterogeneity of evasion. But the stark differences in compliance rates across taxable items that line up closely with detection rates suggest strongly that deterrence is a powerful factor in evasion decisions. Given the current state of theory and evidence on tax evasion, it is not clear in what way or how much enforcement might be most efficiently increased. Although the normative theory of taxation has been extended to tax system instruments such as the intensity of enforcement, the empirical knowledge for operationalizing these rules is sparse.

Bibliography

IRS (Internal Revenue Service). 2006. Updated estimates of the TY 2001 individual income tax underreporting gap. Overview. 22 February. Washington, DC: Office of Research, Analysis, and Statistics, U.S. Department of the Treasury.

IRS Oversight Board. 2006. 2005 Taxpayer attitude survey. U.S. Department of the Treasury. Online. Available at http://www.ustreas.gov/irsob/releases/2006/02212006.pdf. Accessed 28 June 2007.

Slemrod, J. 2007. Cheating ourselves. *Journal of Economic Perspectives* 21(1): 25–48.

Slemrod, J., and J. Bakija. 2004. *Taxing ourselves: A citizen's guide to the debate over taxes*. 3rd ed. Cambridge, MA: MIT Press.

Slemrod, J., and Yitzhaki, S. 2002. Tax avoidance, evasion and administration. In *Handbook of public economics*, vol. 3, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.

Tax Expenditures

Daniel N. Shaviro

Abstract

Labelling certain provisions in the tax law as tax expenditures has been criticized for lacking an 'agreed conceptual model' for distinguishing between integral tax rules and interpolations reflecting spending rather than tax policy. However, the tax expenditure concept can be reformulated as relying on a distinction between (a) the distributional goals that might underlie the use of a tax base such as income or consumption, and (b) allocative goals such as encouraging particular activities or investments. Tax expenditure estimates could be prepared using both measures, including negative tax expenditures (that is, tax penalties) as well as positive ones.

Keywords

Consumption tax base; Consumption taxation; Haig–Simons income taxation; Income tax base; Tax expenditures; Tax penalties; Taxation of corporate profits; Taxation of income

JEL Classifications

H2

The practice of labelling certain provisions in the tax law as 'tax expenditures' is widely attributed to Stanley Surrey, the longtime Harvard law professor and, from 1961 to 1969, Assistant Secretary of the Treasury for Tax Policy in the United States. Surrey introduced the term in a 1967 speech, in which he urged official measurement of the revenue cost of all tax expenditures, which he defined as 'special' benefits in the income tax law. Surrey argued that publication of a tax expenditure budget would encourage and facilitate treating special tax rules on a par with similarly motivated direct spending rules.

This proposal had earlier antecedents, having been part of the federal budgetary process in Germany since at least 1959. In the United States, however, it proved more controversial, reflecting Surrey's use of it as a tool, not just of budgetary policy, but also in tax policy debate, where he was well-known as an advocate of progressive, comprehensive income taxation. In keeping with his tax policy views, Surrey, after leaving the Treasury, pressed the argument that tax expenditures

should generally be eliminated from the US income tax, with any that served meritorious social goals being replaced by direct appropriations. Surrey's advocacy may have encouraged some with different tax policy views to regard the tax expenditure budget as special pleading for his views, merely masquerading as anodyne budgetary reporting.

Concern that the tax expenditure budget unduly served Surrey's particular views became more widespread with the rise in tax policy circles, beginning in the 1970s, of support for replacing the US federal income tax with a consumption tax. Various tax expenditures from an income tax standpoint, such as the exclusion of interest from bonds issued by state or local governments, are correct from a consumption tax standpoint. If 'tax expenditures' should be eliminated presumptively, then using a normatively controversial income tax standpoint to define them could be viewed as unduly aiding those who favour moving the current 'hybrid' US system closer to the income tax pole rather than to the consumption tax pole.

In the face of these criticisms, Surrey arguably won the battle concerning tax expenditure analysis, but lost the war. The Congressional Budget and Impoundment Control Act of 1974 made tax expenditure estimates mandatory both in the President's annual budget and in certain reports by Congressional committees. These estimates generally are static, measuring the level of utilization of a given provision, rather than how much revenue would be raised by repealing it. The tax expenditure concept has remained intellectually controversial, however. Moreover, it has not noticeably discouraged the use of 'special' tax benefits, other than perhaps temporarily if it helped to inspire the landmark Tax Reform Act of 1986.

What Is a Tax Expenditure?

In official US estimates, tax expenditures are defined as pro-taxpayer departures from a 'normal' income tax base. This is not uncommon, although practices vary around the world. The normal income tax base has a number of features that depart from theoretically pure Haig–Simons income taxation. For example, it features a realization requirement under which unrealized gains and losses have no immediate tax consequences, includes double taxation of corporate income, and makes no adjustments for inflation. It also treats as tax expenditures some items whose preferentiality is controversial – for example, the itemized deductions for medical expenses and for state and local income taxes paid.

Criticism of tax expenditure analysis has focused both on what some view as the arbitrariness of the normal income tax base in any of its currently used variants, and on the lack of any 'agreed conceptual model' (Bittker 1969, p. 258) for identifying special tax benefits. Such a model is needed to support the view that a particular provision, although located in the tax code, is actually a spending rule.

A deeper problem is that 'taxes' and 'spending' cannot meaningfully be distinguished, even though the former involves cash flow to the government while the latter involves cash flow from the government. By way of illustration, consider US federal income taxation of Social Security benefits, which was introduced during the Reagan administration and increased during the Clinton administration. While both administrations classified the changes as benefit cuts, Republican critics of the Clinton proposal argued that it was a tax increase. Arguably, these critics were formally correct, in that the reduction in net benefits was accomplished via income tax payments. However, if exactly the same reduction in net benefits had been accomplished by reducing gross benefits (that is, the amounts paid out by the Social Security administration), it evidently would have been a 'spending cut'. Tax expenditure analysis requires discerning a substance to distinction between taxes and spending that does not depend in this way on form, or else by definition everything in the tax code would be a tax rule.

Still, the intuition underlying tax expenditure analysis is hard to dismiss. Suppose, for example, that the US Congress decided to replace $1 billion of military spending with a $1 billion tax credit,

offered to the same taxpayers who would have received the direct appropriation in exchange for the same goods or services. Nominally, this switch would lower both 'taxes' and 'spending' by $1 billion. In substance, however, little would have changed. Tax expenditure analysis would treat the credit as 'spending' through the tax code, thus preventing the change in form from being misperceived as a change in substance.

Distribution and Allocation

The distinction that tax expenditure analysis draws between tax and spending rules can be restated in terms of Richard Musgrave's (1959, p. 5) conceptual division between the public sector's distribution and allocation functions. Apportioning the burden of paying for government through a measure of ability to pay, such as income or consumption, is conceptually a distributional enterprise. Thus, a rule within the income tax law, such as the hypothetical military suppliers' credit, that appears to serve primarily allocative purposes (furnishing goods and services for military use) can logically be viewed as extraneous to the distribution function, even if its placement in the tax code is desirable (for example, on administrative grounds). The same reasoning applies in reverse if a set of tax rules serving primarily allocative purposes includes provisions that appear to serve primarily distributional aims. Thus, suppose a Pigouvian pollution tax offered rebates to low-income polluters. One could extend this reasoning to cover clearly distinguishable allocative or distributional functions as well – for example, the inclusion of education subsidies, such as lower tax rates for pollution by schools, in a Pigouvian pollution tax.

In each of these cases, the extraneous provision could be termed a tax expenditure, albeit without any necessary implication that it should be eliminated or moved. The reason for this linguistic exercise might be to increase public understanding of the provision at issue, and in particular to prevent 'tax cuts' from being distinguished from 'spending increases' on purely formal grounds where their substance is identical.

While this restatement of the distinction that Surrey attempted to draw between taxes and spending can go a long way to rationalize tax expenditure analysis, it does not support all aspects of current practice. For example, in the USA child tax credits are classified as tax expenditures, but personal exemptions (deductions for dependents, including children) are not. Yet the two provisions have similar effects, and both could be viewed as relating to a distributional goal of having tax burdens depend on family size. Thus, neither the distinction between them nor the treatment of either as a tax expenditure is highly robust.

Income Tax Base Versus Consumption Tax Base

The distinction between distribution and allocation functions does not address the issue of how to handle distinctions between the income and consumption tax bases, as illustrated by the question of whether the US exclusion for state and local government bond interest is a tax expenditure. Here the answer would depend on whether the distribution branch was assumed to follow income or consumption tax norms. Under the latter norm, the anomalous result would be taxing other interest income, rather than exempting municipal bond interest. From either distributional perspective, however, the distinction in the tax treatment for interest depending on who paid it is likely to seem anomalous, even if desirable for allocative reasons. One possible solution discussed in recent US government budgets is to prepare alternative tax expenditure listings, one from an income tax baseline and the other from a consumption tax baseline.

Administrative Departures from a Pure Income Tax or Consumption Tax

One reason for the practice of computing tax expenditures relative to a 'normal' income tax base, rather than Haig–Simons income, is the notion that the provisions being analysed are substitutes for direct spending. Thus, if the main reason for not taxing unrealized asset appreciation

is administrative, the legislature is unlikely to be choosing between alternative implementations of the resulting allocative policy. However, if administrative considerations limit the departures from a given theoretical base that are treated as tax expenditures, a system with those departures may easily be confused with one that actually implemented the theoretical ideal. One possible response to this dilemma is to create a separate category in tax expenditure analysis for departures from a given theoretical base that appear to be primarily administratively motivated (Shaviro 2004, p. 218).

Negative Tax Expenditures (Tax Penalties)

Under current practice, tax expenditure analysis is limited to measuring the static revenue effect of departures from a given baseline that favour the taxpayer. Departures that disfavour a taxpayer are ignored, rather than being treated as negative tax expenditures or tax penalties. However, the rationale for measuring departures in one direction arguably should apply symmetrically.

An example of an unmeasured tax penalty in the current US income tax is the disallowance of business expense deductions for bribes. However socially desirable the disallowance rule may be, it reflects a departure from simply measuring net income in cases where the bribe was economically motivated. The practical importance of measuring tax penalties would increase if income tax rules were being analysed from a consumption tax as well as an income tax baseline, since this would cause a variety of common income tax features, such as taxing particular kinds of interest income, to constitute tax penalties.

Bibliography

Bittker, B. 1969. Accounting for federal 'tax subsidies' in the national budget. *National Tax Journal* 22: 244–261.

Musgrave, R. 1959. *The theory of public finance*. New York: McGraw Hill.

Shaviro, D. 2004. Rethinking tax expenditures and fiscal language. *Tax Law Review* 57: 187–232.

Surrey, S. 1973. *Pathways to tax reform*. Cambridge, MA: Harvard University Press.

Surrey, S., and P. McDaniel. 1985. *Tax expenditures*. Cambridge, MA: Harvard University Press.

United States Government. 2003. Tax expenditures. Ch. 6 in fiscal year 2003. *Budget of the United States government.* Analytical perspectives. Online. Available at http://www.gpoaccess.gov/usbudget/fy03/pdf/spec.pdf. Accessed 9 Feb 2007.

Tax Havens

James R. Hines Jr.

Abstract

Tax havens are low-tax jurisdictions that offer businesses and individuals opportunities for tax avoidance. The 45 major tax haven countries in the world today are small, affluent, and generally well governed. They attract disproportionate shares of world foreign direct investment, and, largely as a consequence, their economies have grown much more rapidly than the world as a whole since the 1980s. The effect of tax havens on economic welfare in high-tax countries is unclear, though the availability of tax havens appears to stimulate economic activity in nearby high-tax countries.

Keywords

Foreign investment; Tax avoidance; Tax havens; Transfer pricing

JEL Classifications

H3

Tax havens are low-tax jurisdictions that offer businesses and individuals opportunities for tax avoidance.

There are roughly 45 major tax havens in the world today. Examples include Andorra, Ireland, Luxembourg and Monaco in Europe, Hong Kong and Singapore in Asia, and the Cayman Islands, the Netherlands Antilles, and Panama in the Americas. These tax havens are generally small

and affluent, in total comprising just 0.8 per cent of world population, though accounting for 2.3 per cent of world income (Hines 2005). Low-tax jurisdictions are also common within countries, at various times taking the form of special economic zones in China, offshore possessions and local enterprise zones in the United States, and tax-favoured regions including eastern Germany, southern Italy, eastern Canada, and others. Tax havens are widely used by international investors; in 1999, 59 per cent of US multinational firms with significant foreign operations had affiliates in one or more tax havens (Desai et al. 2006b).

Tax Haven Experiences

Countries offer low tax rates in the belief that, by doing so, they attract greater investment and economic activity than would otherwise have been forthcoming. Countries with low tax rates permit investors to retain most of their locally earned pre-tax income; other considerations equal, therefore, countries with lower tax rates should be expected to offer a broader range of attractive opportunities, and therefore draw larger volumes of foreign investment, than countries with higher tax rates.

The possibility of using tax havens to facilitate avoidance of taxes that would otherwise be owed to governments of other countries adds to the attractiveness of tax haven investments. For individuals, who are taxed by their home governments on income earned in tax havens, tax avoidance typically entails wilful income misreporting. For businesses, tax avoidance can be accomplished by the use of financial arrangements, such as intrafirm lending, that locate taxable income in low- tax jurisdictions and tax deductions in high-tax jurisdictions. In addition, firms are often able to adjust the prices at which affiliates located in different countries sell goods and services to each other. Most governments require that firms use arm's-length prices, those that would be used by unrelated parties transacting at arm's length, for transactions between related parties, in principle thereby limiting the scope of tax-motivated transfer price adjustments. In practice, however, the indeterminacy of appropriate arm's length prices for many goods and services, particularly those that are intangible, or for which comparable unrelated transactions are difficult to find, leaves room for considerable discretion. As a result, transactions with tax haven affiliates can be used to reallocate income from high-tax locations to the tax haven affiliates themselves or else to other low-tax foreign locations. This, in turn, increases the appeal of locating investment in foreign tax havens.

As a result of these incentives, American firms exhibit unusual activity levels and income production in foreign tax havens (Hines 2005). Of the property, plant and equipment held abroad by American firms in 1999, 8.4 per cent was located in tax havens, considerably more than would be predicted strictly on the basis of the sizes of tax haven economies. Employment abroad by American firms was likewise unusually concentrated in foreign tax havens, with 6.1 of total foreign employee compensation, and 5.7 per cent of total foreign employment, located in tax haven affiliates. American firms located 15.7 per cent of their gross foreign assets in the major tax havens in 1999; the major foreign tax havens accounted for 13.4 per cent of their total foreign sales, and a staggering 30 per cent of total foreign income in 1999. Much reported tax haven income consists of financial flows from other foreign affiliates that parent companies owned indirectly through their tax haven affiliates.

Tax haven countries have enjoyed very rapid economic growth rates that coincide with dramatic inflows of foreign investment. Tax havens averaged 3.3 per cent annual per capita real GDP growth from 1982 to 1999, whereas the world averaged just 1.4 per cent annual real per capita GDP growth over the same period. Controlling for country size, initial wealth, and other observable variables, does not change the conclusion that the period of globalization has been favourable for the economies of countries with very low tax rates (Hines 2005).

The policy of offering foreigners very low tax rates is potentially costly to tax haven

governments, if doing so reduces tax collections that might otherwise have been used to fund worthwhile government expenditures. It is far from clear, however, that tax haven countries face significant trade-offs of this nature. Governments have at their disposal many tax instruments, including personal income taxes, property taxes, consumption or sales taxes, excise taxes, and others, that can be used to finance expenditures. Furthermore, even very low rates of direct taxation of business investment may yield significant tax revenues if economic activity expands in response. In fact, the public sectors of tax haven countries are of comparable sizes to those of other countries, though there is evidence that they may be somewhat smaller than would have been predicted on the basis of their populations and affluence alone (Hines 2005).

Characteristics of Tax Havens

Tax havens are small countries, commonly below one million in population, and are generally more affluent than other countries. In addition, tax havens score very well on cross-country measures of governance quality that include measures of voice and accountability, political stability, government effectiveness, rule of law, and control of corruption. Indeed, there are almost no poorly governed tax havens. Poorly governed countries, of which the world has many, almost never become tax havens (Dharmapala and Hines 2006).

An important reason why better-governed countries are more likely than others to become tax havens is that the potential returns are greater: higher foreign investment flows, and the economic benefits that accompany them, are more likely to accompany tax reductions in well-governed countries than they are tax reductions in poorly governed countries. Evidence from the behaviour of American firms is consistent with this explanation, in that tax rate differences among well-governed countries are associated with much larger effects on US investment levels than are tax rate differences among poorly governed countries (Dharmapala and Hines 2006).

Impact on Other Countries

Tax havens are viewed with alarm in parts of the high-tax world, where there are concerns that the availability of foreign tax haven locations may divert economic activity from countries with higher tax rates, and erode their tax bases. Alternatively, tax havens could encourage investment in other countries, if the ability to relocate taxable income into tax havens improves the desirability of investing in high-tax locations, or if low tax rates reduce the cost of goods and services that are inputs to production or sales in high-tax countries. In fact, the evidence indicates that foreign tax haven activity appears to stimulate activity in nearby high-tax countries, a one per cent greater likelihood of establishing a tax haven affiliate being associated with two-thirds of a per cent greater investment and sales in nearby non-haven countries (Desai et al. 2006a).

The empirical regularity that tax havens stimulate economic activity in high-tax countries does not resolve the impact of tax havens on the welfare of high-tax countries. Tax avoidance carries mixed implications for governments of nearby countries, since it may erode tax bases and therefore tax collections, implying that the greater economic activity associated with nearby tax havens might come at a high cost in terms of forgone government revenues. One possibility is that countries would prefer to subject mobile international companies to lower tax rates than they do other firms, but are prevented from doing so by political considerations or the practical difficulty of distinguishing multinational from domestic firms. In such a setting, countries might benefit from permitting multinational firms to obtain tax reductions by using affiliates in tax havens, thereby implicitly subjecting these mobile firms to lower tax burdens than other taxpayers.

In 1998, the Organization for Economic Co-operation and Development (OECD) introduced its Harmful Tax Practices initiative, the purpose of which was to discourage OECD member countries and certain tax havens from pursuing policies that were thought to harm other countries by unfairly eroding tax bases. As part

of this initiative, the OECD produced a List of Un-Cooperative Tax Havens, identifying countries that have not committed to sufficient exchange of information with tax authorities in other countries. The concern was that the absence of information exchange might impede the ability of OECD and other countries to tax their resident individuals and corporations on income or assets hidden in foreign tax havens. As a result of the OECD initiative, along with diplomatic and other actions of individual nations, many countries and jurisdictions outside the OECD have committed to improve the transparency of their tax systems and to facilitate information exchange. While there remain a few tax havens that have not made such commitments, the vast majority of the world's tax havens rely on low tax rates and other favourable tax provisions to attract investment, rather than using the prospect of local transactions that will not be reported.

See Also

- ▶ Tax Competition
- ▶ Tax Treaties
- ▶ Taxation of Foreign Income
- ▶ Transfer Pricing

Bibliography

Desai, M.A., C.F. Foley, and J.R. Hines Jr. 2006a. Do tax havens divert economic activity? *Economics Letters* 90: 219–224.

Desai, M.A., C.F. Foley, and J.R. Hines Jr. 2006b. The demand for tax haven operations. *Journal of Public Economics* 90: 513–531.

Dharmapala, D., and J.R. Hines, Jr. 2006. *Which countries become tax havens?*, Working paper No. 12802. Cambridge, MA: NBER.

Hines, J.R., Jr. 2005. Do tax havens flourish? In *Tax policy and the economy,* vol. 19, ed. J.M. Poterba. Cambridge, MA: MIT Press.

Hines Jr., J.R., and E.M. Rice. 1994. Fiscal paradise: Foreign tax havens and American business. *Quarterly Journal of Economics* 109: 149–182.

Rose, A.K., and M.M. Spiegel. 2007. Offshore financial centers: Parasites or symbionts? *Economic Journal* 117: 1310–1335.

Slemrod, J., and J.D. Wilson. 2006. *Tax competition with parasitic tax havens*, Working paper No. 12225. Cambridge, MA: NBER.

Tax Incidence

Gilbert E. Metcalf

Abstract

Tax incidence is the study of who bears the burden of a tax. It distinguishes between statutory incidence (the legal requirement to remit a tax) and economic incidence (the change in real income or wealth resulting from a tax). Considerable advances have been made since the mid-1980s in our understanding of the burden of taxes in imperfectly competitive models as well as in intertemporal models. In particular, analysing lifetime tax burdens can give markedly different results for many taxes. Increases in computing power and the availability of large-scale data-sets have also enriched our understanding of tax incidence.

Keywords

Ad valorem taxes; Commodity taxes; Consumption tax; Excise taxes; Factor taxes; Imperfect competition; Lifetime income; Lump-sum taxes; Progressive and regressive taxation; Property taxation; Statutory and economic tax incidence; Tax incidence; Taxation of corporate profits

JEL Classifications

H2

Tax incidence is the study of who bears the economic burden of a tax. Broadly put, it is the positive analysis of the impact of taxes on the distribution of welfare within a society. It begins with the very basic insight that the person who has the legal obligation to pay a tax may not be the person whose welfare is reduced by the tax. The statutory incidence of a tax refers to the distribution of tax payments based on the legal obligation to remit taxes to the government. Thus, for example, the statutory burden of the payroll tax in the United States is shared equally between employers and

employees. Economists, quite rightly, focus on the economic incidence, which measures the changes in economic welfare in society arising from a tax. The standard view of the economic burden of the payroll tax in the United States is that it is borne entirely by employees.

Economic incidence differs from statutory incidence because of changes in behaviour and consequent changes in equilibrium prices. Consumers buy less of a taxed product, so firms produce less and buy fewer inputs – which changes the net price of each input. Thus the job of the incidence analyst is to determine how those other prices change, and how those changes affect different kinds of individuals.

The distributional impact of a tax (or system of taxes) depends in part on how the question is framed. An absolute incidence analysis considers the burden of a change in taxes without regard to the use of proceeds. A differential incidence analysis carries out a revenue-neutral change in tax by raising one tax while lowering another. Typically, a lump-sum tax is changed to effect revenue neutrality. A balanced budget incidence analysis considers the burden of a change in taxes along with an equivalent change in spending. In his classic analysis of the US tax system, Pechman (1985) carried out a differential incidence analysis and concluded that the total system of taxes in the United States was broadly proportional. Taking into account government transfers financed by taxes, on the other hand, Browning and Johnson (1979) argued that the US tax system was progressive.

In addition to framing the incidence question precisely, incidence results can depend on the time frame for analysis. Pechman's analysis ranks households by their annual income. It is well known that annual income can be a poor proxy for measuring the well-being and consumption potential of a household, because of measurement error and lifetime income considerations. Lifetime income considerations are particularly important for assessing the distributional implications of a consumption tax, since consumption to annual income ratios are very high in the lowest annual-income deciles. Fullerton and Rogers (1993) replicate the Pechman analysis using a lifetime income framework, and conclude that the overall incidence of the US tax system is similar to that obtained in Pechman's annual income framework, though the forces driving incidence results differ somewhat.

In a perfectly competitive partial equilibrium framework, the economic incidence of a tax is unaffected by which side of the market the tax is levied on. Thus the statutory requirement to share the payroll tax in the United States equally between employer and employee has no bearing on the ultimate incidence of the tax. Second, the economic burden of a tax is borne more heavily by the side of the market that is less elastic (in absolute value). Thus, the share of the payroll tax borne by the employee is, to a first-order approximation, equal to $\varepsilon_D/(\varepsilon_s + \varepsilon_D)$ where ε_S (ε_D) is the labour supply (demand) elasticity.

This burden share formula suggests that no more than 100 per cent of the tax can be shifted to a party. In an imperfectly competitive market, commodity tax overshifting can occur (in the sense that the consumer price rises by more than the tax rate). Moreover, *ad valorem* and excise taxes, which have equivalent burden impacts in a competitive market when set to collect the same revenue, now can have different burden impacts. Delipalla and Keen (1992) show that in markets with oligopoly supply *ad valorem* taxes are less likely to lead to overshifting than excise taxes. Once one allows for imperfect competition, many counter-intuitive results can obtain, including a commodity tax *reducing* the consumer price (for example, Cremer and Thisse 1994). Fullerton and Metcalf (2002) develop the analysis of tax incidence under imperfect competition further and provide some hypothetical results.

Harberger (1962) is the progenitor of the modern field of general equilibrium incidence analysis. In addition to providing a framework for analysing the corporate income tax, Harberger's approach can be used to analyse a wide array of taxes. He models the corporate income tax as a partial factor tax, that is, a tax on the use of one factor in one sector. The tax thus affects relative factor prices and relative output prices. Harberger concludes that capital is likely to bear approximately the full burden of the corporate income tax.

Capital mobility means that the burden is on all capital, not just corporate capital.

Harberger's analysis assumed a closed economy. In a small open economy with international capital mobility, corporate tax drives capital abroad so that domestic savers earn the same net return as before the tax is imposed. This drives down the domestic capital stock, and thus the domestic wage rate, and the burden of the tax falls on labour (as an immobile factor). While the immobile local factor bears a burden from the tax, Bradford (1978) shows that worldwide capital in the aggregate suffers a loss exactly offset by gains to immobile factors in the rest of the world, resulting from the outflow of capital from the country imposing the tax. In contrast, Gravelle and Smetters (2001) argue that imperfect substitutability of domestic and foreign products can limit or even eliminate the incidence borne by labour, even in an open economy model. They find that the tax is borne by domestic capital, as in the original Harberger model.

While Harberger's analysis (and subsequent work) showed the importance of general equilibrium effects, it lacked a fully dynamic characterization of savings and investment, channels through which important burden shifting could occur. Feldstein (1974), for example, argues that much (if not all) of the burden of a tax on capital income is shifted to workers in the form of lower wages as a result of decreased investment reducing the capital–labour ratio.

Once investment is considered, the incidence of a tax in a dynamic model can also be affected by the distinction between old and new capital. Old capital is capital in place prior to a tax change. For example, Auerbach and Kotlikoff (1987) show that a consumption tax and a wage tax – two approaches to exempting capital income from taxation – differ only in their tax treatment of old capital. In the absence of transition rules, the former subjects old capital to a lump-sum tax, while the latter does not. In addition to distributional implications, the presence of old capital complicates the attribution of economic incidence. Consider a new property tax that has been in place for many years in a community. Carrying out an incidence analysis today, we might allocate the burden of the tax to current owners based on their property values. This approach would be consistent with the 'old' view of property taxation (see Fullerton and Metcalf 2002, for more on the incidence of the property tax). But with capitalization effects the tax burden should properly be allocated to the property tax owners at the time of the enactment of the tax: more precisely, it would be allocated to the owners at the time that potential buyers and sellers of property in the community become aware that the tax would be enacted. Without offsetting benefits from the property tax revenues, potential homeowners will be willing to pay less for housing. In equilibrium, housing values would fall by the present discounted value of the stream of future tax payments at the time of enactment, and the owners at that time would bear the entire burden of the tax.

To return to the corporate income tax, an increase in the tax rate generates lump-sum taxes on previously installed capital through capitalization effects. As Auerbach (2005) emphasizes, the tax treatment of corporate capital is sufficiently complicated to ensure that assigning its burden is a hazardous exercise, but both in the short and the long run it is probably the case that some portion of the tax falls specifically on shareholders due to the tax on old capital, among other factors.

Careful tax incidence analysis is essential to understanding the distributional implications of a country's tax system. The field of incidence analysis has progressed dramatically since the mid-1980s, as new research has yielded fresh insights into the burden of taxes in imperfectly competitive models and in intertemporal models. The increase in computing power and the availability of large-scale data-sets have also enriched our understanding of tax incidence. Despite all the advances that have occurred, the topic of tax incidence will probably continue to be an area of productive research, yielding further insights in the years to come.

See Also

► Income Taxation and Optimal Policies
► Progressive and Regressive Taxation

- ▶ Public Finance
- ▶ Redistribution of Income and Wealth
- ▶ Taxation of Corporate Profits

Bibliography

Auerbach, A. 2005. Who bears the corporate tax? A review of what we know. Working Paper No. 11686. Cambridge, MA: NBER.

Auerbach, A.J., and L.J. Kotlikoff. 1987. *Dynamic fiscal policy*. New York: Cambridge University Press.

Bradford, D.F. 1978. Factor prices may be constant but factor returns are not. *Economics Letters* 1: 199–203.

Browning, E.K., and W.R. Johnson. 1979. *The distribution of the tax burden*. Washington, DC: American Enterprise Institute.

Cremer, H., and J.F. Thisse. 1994. Commodity taxation in a differentiated oligopoly. *International Economic Review* 35: 613–633.

Delipalla, S., and M. Keen. 1992. The comparison between ad valorem and specific taxation under imperfect competition. *Journal of Public Economics* 49: 351–367.

Feldstein, M. 1974. Incidence of a capital income tax in a growing economy with variable savings rates. *Review of Economic Studies* 41: 505–513.

Fullerton, D., and G.E. Metcalf. 2002. Tax incidence. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, vol. 4. Amsterdam: North-Holland.

Fullerton, D., and D.L. Rogers. 1993. *Who bears the lifetime tax burden?* Washington, DC: Brookings Institution.

Gravelle, J.G., Smetters, K. 2001. Who bears the burden of the corporate tax in the open economy? Working Paper No. 8280. Cambridge, MA: NBER.

Harberger, A.C. 1962. The incidence of the corporation income tax. *Journal of Political Economy* 70: 215–240.

Pechman, J. 1985. *Who paid the taxes: 1966–85?* Washington, DC: Brookings Institution Press.

Tax Shelters

David A. Weisbach

Abstract

Tax shelters take advantage of unintentional gaps in the tax base often caused by subtle mismatches in complex tax rules. The optimal line between allowable tax planning and illegitimate tax shelters depends on the cost of closing these gaps compared with the revenue raised, relative to the efficiency costs of other sources of funds. The definition of illegitimate tax shelters, therefore, depends on parameters such as the tax base and rate structure as well as the expected taxpayer response to different possible definitions.

Keywords

Elasticity of taxable income; Tax avoidance; Tax base; Tax compliance; Tax evasion; Tax rules; Tax shelters

JEL Classifications

H2

The term 'tax shelters' generally refers to any tax reducing activity other than outright evasion or traditionally modelled responses to taxation, such as changes in labour supply or savings. The term sometimes includes investing in explicitly tax-favoured assets such as homes, life insurance or tax exempt bonds. At other times, however, the term is used to mean only tax-reducing activities inconsistent with the intent of the tax law, in which case it may not include these activities. The concept is closely related to, but usually thought to be narrower than, the notion of tax avoidance.

To define the term more precisely has proven to be impossible. The US Treasury Department (1999) observed that tax shelters come in the 'guises of Proteus' and argued that no single definition was appropriate. The tax law itself defines shelters for purposes of certain penalties as a plan or arrangement, a significant purpose of which is the avoidance or evasion of federal income tax, a definition sufficiently broad as to be almost meaningless. Rather than attempt to define shelters, Treasury (1999) has listed factors common to many shelter transactions, including (1) lack of economic substance; (2) inconsistent financial and accounting treatment; (3) presence of tax-indifferent parties; (4) complexity; (5) unnecessary steps or novel investments; (6) promotion or marketing; (7) confidentiality; (8) high transaction costs; and (9) risk reduction arrangements. The Joint Committee on Taxation (1999) took a similar approach, recommending the use of tax

shelter indicators or factors for purposes of triggering enhanced disclosure requirements or penalties. Other investigations have defined tax shelters as complex transactions, marketed by sophisticated professionals, used by corporations or wealthy individuals to obtain significant tax benefits in a manner never intended by the tax law (Levin 2003; GAO 2003). Perhaps the most pithy definition, by Michael Graetz, is that a tax shelter is 'a deal done by very smart people that absent tax considerations, would be very stupid' (Herman 1999).

The definition of tax shelters matters for two related reasons. First, it points to behavioural responses to taxation that are left out of the usual analysis of labour/ leisure or investment distortions and that are also not included in many analyses of tax evasion. Tax evasion, for example, is usually modelled as a report/non-report decision that imposes risk on the individual but otherwise has no direct affect on behaviour. Tax shelters, in contrast to evasion, normally involve actual although subtle changes to behaviour, such as leasing rather than owning, using a different organizational form, or using hybrid financing instruments. A less than optimal use of legal forms can produce social losses beyond merely the risk of audit. Analyses of investment distortions normally compute marginal effective tax rates on different activities. Tax shelters can significantly reduce effective tax rates on certain activities, which means that the distortions might be substantially larger than otherwise thought. Second, the definition has a set of legal consequences, such as disallowance of tax benefits, penalties, disclosure rules, and additional audits.

The appropriate legal consequences of the definition of tax shelters and an analysis of the economic effects of sheltering need to be tied together. In particular, the scope of activities that should be subject to various policies must be determined based on an analysis of the consequences of such policies, not a definition. There is no clear line between various tax-reducing activities, ranging from working less, being paid in tax-free fringe benefits, investing in tax-favoured assets, entering into traditional shelters, and false reporting. The treatment of a class of these activities as tax shelters, others as criminal evasion, and others as allowable should depend on which activities are optimally subject to a given set of policy instruments.

This approach means that, to define shelters, we need a general theory of tax instruments, including the tax base, the penalty structure, the drafting of legal rules, reporting regimes, and the audit rate. As emphasized by Feldstein (1995, 1999), given some tax base and set of audit, penalty, and similar policies, the private marginal cost of all tax reducing activities will be equal. Therefore, the elasticity of taxable income can be used as a summary measure of the efficiency of the tax system. Slemrod and Kopczuk (2002) observe that the elasticity of taxable income is, in part, a policy variable rather than a preference because policymakers can control the size of the tax base, auditing mechanisms, penalties, and other variables that affect opportunities to reduce taxable income. First order conditions for an arbitrary tax instrument (assuming an optimal linear income tax) set the marginal administrative cost of the instrument equal to sum of marginal (indirect) utility from the change in the instrument and the marginal revenue. Marginal revenue is made up of two components: revenue directly from the change in taxation of the item at issue with the elasticity held constant, and revenue from the change in the elasticity of taxable income.

Viewed in this way, tax shelters are similar to any other gap in the tax base, and appropriate responses to shelters (and definition of shelters) depend on relative cost of obtaining that source of funds. Although this general formulation does not tell us anything about the particular definition of tax shelters or which particular mix of instruments is optimal, it does focus attention on the relevant factors. For example, it seems clear that there should be substantial sanctions for fraud because any other rule would produce a very high elasticity of taxable income – without such sanctions, any increase in the tax rate, starting from zero, would produce substantial reductions in reported income. A similar conclusion holds for many shelters. If they became well known and inexpensive, the elasticity of taxable income becomes unduly high. Similarly, an important effect of

imposing a tax on a shelter is reducing the elasticity of taxable income rather than raising revenue directly from the tax on the shelter activity. Because few would engage in the shelter activity without the tax benefits, any revenue from a direct tax on that activity is likely to be small. Most important, the elasticity measure emphasizes that the optimal definition of tax shelters depends on what other instruments are in use, such as the scope of the tax base, the rates, and the audit and evasion rules. For example, *ceteris paribus*, the higher the tax rates, the broader the optimal definition of shelters is likely to be.

Another approach, more grounded in law than in economics, is to focus on the drafting of tax rules. The primary cause of shelters, in this view, is the imperfect interactions of statutory rules. Treasury (1999) referred to these interactions as discontinuities. Given limited resources, drafters of tax rules can cover only general cases. Taxpayers have a private incentive to find unusual interactions of otherwise reasonable rules and structure transactions to take advantage of them. Weisbach (1999) argued that the solution to this problem, long followed by US law, involves general rules for common behaviours and ambiguous standards, so-called anti-abuse rules, that prevent intentional use of unintended, tax-reducing interactions in the general rules. This approach balances the uncertainty of the ambiguous standards (and the potential principal–agent problems of a revenue-maximizing tax agency overusing the standards) with the benefit of tax rules that need to cover only general cases and, therefore, that can be simpler.

A related approach is to focus on the industrial organization of the tax shelter industry. The designing and implementation of tax shelters takes significant resources. While privately beneficial, most of this expenditure of resources has no social benefit. On the other hand, advice about compliance (as opposed to shelters) does have a social benefit. The regulatory problem is to distinguish these activities. Various approaches have been considered, such as limiting the use of contingency fee arrangements based on tax savings, requiring disclosure by advisors of client lists, and direct sanctions, including criminal penalties, for giving inappropriate tax advice.

Reliable estimates of the number or size of tax shelters are difficult to obtain because of their secrecy and because of the ambiguity in the definition of shelters. There is no tax shelter equivalent to the measurements of the tax gap (which is a measure of non-compliance). Shelters have long been considered a problem in the tax law, and anecdotal evidence of sheltering activity has frequently driven tax policy. Important Supreme Court decisions on shelters date back to the 1930s or earlier. In 1934, the Treasury Department attempted to prosecute the former Secretary of the Treasury, Andrew Mellon, for tax evasion. Although the grand jury refused to indict, Mellon was eventually ordered to pay $400,000 in back taxes for what might be considered shelters. As Secretary, Mellon had solicited from Internal Revenue Service 'memorandum setting forth the various ways by which an individual may legally avoid tax' (Brownlee 1996). In 1969, an outgoing Secretary of the Treasury revealed that in 1967 no income taxes were paid on 155 tax returns with gross incomes of 200,000 dollars or more, including 21 of returns of millionaires (Zelizer 1998). This revelation led immediately to a variety of tax law changes to prevent the use of shelters.

The late 1970s and early 1980s saw a significant rise in tax shelter activity. In 1980, the Commissioner of the Internal Revenue Service stated that 'about 200,000 individual tax returns representing about 18,000 shelter schemes are now at various stages of the examinations' process' (Kurtz 1980). In 1985, there were over 20,000 tax shelter cases pending in the tax court, and, as of 1982, these cases made up approximately one-third of the court's docket (Collinson 1985). Limitations on tax shelter losses that were part of the Tax Reform Act of 1986 were estimated to raise almost $53 billion over the five-year revenue window. Birnbaum and Murray (1987) report that these provisions were central to the compromise that allowed the 1986 Tax Reform Act to pass.

Notwithstanding the tax shelter limitations enacted in the 1986 Act, tax shelter activity was thought to rebound significantly in the 1990s. A Senate Subcommittee investigation reported that a single major accounting firm had more than 500 tax shelter products in its inventory and

that it sold these products aggressively to individuals and corporations (Levin 2003). Some 19 lawyers or accountants associated with these shelters were indicted in the largest criminal tax case in US history. Graham and Tucker (2006) study 44 instances of tax shelters by examining reported cases (which should represent only a small fraction of actual shelters). They find that typical tax shelter deduction was more than one billion dollars per firm and that the median shelter produced a deduction sufficient to shield income of approximately nine per cent of asset value. Settlements from a single type of shelter, known as 'Son of Boss' sold largely to individuals produced more than $3.7 billion in additional taxes in 2005.

See Also

- ▶ Excess Burden of Taxation
- ▶ Tax Compliance and Tax Evasion
- ▶ Tax Havens

Bibliography

Birnbaum, J.H., and A.S. Murray. 1987. *Showdown at Gucci Gulch*. New York: Random House.

Brownlee, W.E. 1996. *Federal taxation in America: A short history*. Cambridge: Cambridge University Press.

Collinson, D. 1985. New York State Bar Association Tax Section Committee on Practice and Procedure, 'Managing the Tax Court Docket'. *Tax Notes Today* 85: 146–193 (24 July).

Desai, M. 2003. The divergence between book and tax income. *Tax Policy and the Economy* 17: 169–206.

Feldstein, M.S. 1995. The effect of marginal tax rates on taxable income: A panel study of the 1986 tax reform act. *Journal of Political Economy* 103: 551–572.

Feldstein, M.S. 1999. Tax avoidance and the deadweight loss of the income tax. *Review of Economics and Statistics* 81: 674–680.

GAO (General Accounting Office). 2003. Internal revenue service: Challenges remain in combating abusive tax shelters. Testimony before the Committee on Finance. GAO-04-104T. Washington, DC: US Senate.

Graham, J.R., and A. Tucker. 2006. Tax shelters and corporate debt policy. *Journal of Financial Economics* 81: 563–594.

Herman, T. 1999. Tax report. *Wall Street Journal*, 10 Feb, p. A-1.

Joint Committee on Taxation. 1999. Study of present-law penalty and interest provisions as required by Section 3801 of the Internal Revenue Service Restructuring and Reform Act of 1998 (including provisions relating to corporate tax shelters). JCS 3-99. Washington, DC: US Government Printing Office.

Kurtz, J. 1980. Kurtz on 'abusive tax shelters'. *Tax Notes* 10: 213.

Levin, C. 2003. *U.S. tax shelter industry: The role of accountants, lawyers, and financial professionals. Report prepared by the Minority Staff of the Permanent Subcommittee on Investigations*. Washington, DC: US Senate.

Manzon, G. Jr., and G. Plesko. 2002. The relation between financial and tax reporting measures of income. *Tax Law Review* 55: 175–214.

Slemrod, J., and W. Kopczuk. 2002. The optimal elasticity of taxable income. *Journal of Public Economics* 84: 91–112.

U.S. Treasury Department. 1999. *The problem of corporate tax shelters: Discussion, analysis, and legislative proposals*. Washington, DC: U.S. Treasury Department.

Weisbach, D. 1999. Formalism in the tax law. *University of Chicago Law Review* 66: 860–886.

Zelizer, J.E. 1998. *Taxing America: Wilbur D. Mills, Congress, and the State, 1945–1975*. Cambridge: Cambridge University Press.

Tax Treaties

Charles E. McLure Jr

Abstract

Tax treaties coordinate how signatories tax transnational income flows to avoid double taxation and prevent fiscal evasion. Treaties affect the division of tax revenues and signal commitment to international 'rules of the game'. Most of the 2,000 plus extant bilateral tax treaties are based on the OECD Model Tax Convention. Residence countries avoid double taxation of foreign-source business income by exempting it or providing credits for source-country taxes. Source-countries typically reduce withholding taxes on interest, dividends and royalties. Despite apparent differences, the economic effects of taxing worldwide income, with credits for foreign taxes, may resemble those of exempting it.

Keywords

Arm's length prices; Business income; Capital export neutrality; Capital import neutrality; Capital taxation; Characterization of income; Division of tax revenues; Double non-taxation; Double taxation; Electronic commerce; Exchange of information; Fiscal evasion; Foreign-source income, taxation of; Formulary apportionment; General Agreement on Tariffs and Trade; Intangible assets; Mutual Agreement Procedures; Non-discrimination; OECD Model Tax Convention; Permanent establishment; Portfolio investment; Residence country; Signalling; Source country; Tax competition; Tax evasion; Tax exemption; Tax harmonization; Tax havens; Tax sparing; Tax treaties; Taxation of foreign income; Transfer prices; Treaty shopping; UN Model Tax Convention; Withholding taxes

JEL Classifications

H8

The term 'tax treaties' is commonly used – as here – to describe bilateral treaties that coordinate how signatories apply their taxes on income and capital to transnational economic activity. (A few multilateral treaties addressing broader objectives, for example, that establishing the European Union, deal to a limited extent with these or other taxes, as do some treaties – notably the multilateral General Agreement on Tariffs and Trade – that deal primarily with other forms of taxation.) The primary objectives of tax treaties are avoidance of double taxation and prevention of fiscal evasion. Secondary objectives include the division of tax revenues between treaty partners and signalling that signatories will abide by international 'rules of the game'. Avoidance of double non-taxation has recently received increased attention. While treaties generally regulate taxation of both individuals and legal entities, the latter are by far the more important and the focus of this article.

Over 2,000 bilateral tax treaties are currently in force. Because tax treaties require several years of negotiation, they are expected to remain in force for several decades, and their wording is rather general, to allow reinterpretation. The vast majority of tax treaties are based on the OECD Model Tax Convention (OECD 2005), whose origins can be traced to the work of the League of Nations during the 1920s (see Graetz and O'Hear 1997). The 'Commentary' that accompanies and interprets this Model Convention, often also called the OECD Model Tax Treaty, is frequently revised to deal with unforeseen issues (for instance, those involving electronic commerce discussed below). Developing countries generally prefer the United Nations (UN) Model Tax Convention, which is more favourable to their interests as source countries. Although they have historically had difficulty getting more powerful developed countries to accept its terms, this has changed recently (see Kosters 2004). Some countries, including the United States, publish their own model treaties. Although based on the OECD Model, these models deal with special concerns or features of the country's tax system, such as the US concern with 'treaty shopping' considered below. The OECD website discusses many issues covered here.

Avoiding Double Taxation

Nations have the legal right to tax both income arising within their borders and the income of their residents, whatever its geographic source. In the absence of treaties, there is a risk of 'double international taxation' by the 'source country' where income arises and the 'residence country' of the taxpayer, even though domestic legislation may unilaterally provide relief from double taxation. Moreover, because of differences in definitions of residence or source, more than one country may impose either a residence-or a source-based tax. Double taxation impedes transnational transactions and investment flows.

Tax treaties commonly assign to source countries the primary right to tax business income resulting from direct investment, but to residence countries the primary right to tax other forms of income, including that from portfolio investment. Treaties generally provide one of two methods that residence countries can use to avoid double

taxation of foreign-source business income, including dividends from subsidiaries: exemption of foreign-source income and credits for taxes paid to source countries. By comparison, they typically provide that source-country withholding taxes on interest, dividends, and royalties will be reduced reciprocally, sometimes to zero. The latter provisions have given rise to 'treaty shopping', the practice of establishing subsidiaries in a country solely for the purpose of benefiting from its treaties. The United States now insists on a 'limitations of benefits' provision that eliminates most treaty shopping.

Treaties specify rules for the characterization of income (for example, as business profits, dividends, interest, royalties, capital gains and service income) and the geographical 'sourcing' of each type of income. The former task is particularly challenging in cases such as income from the provision of computer software, which might reasonably be characterized as business profits from the sale of goods, royalties or service income.

Treaties limit source-country taxation of business profits to income earned by a permanent establishment (PE), indicated by the presence of a fixed place of business or a dependent agent in the country. This provision has important implications for the division of revenues between source and residence countries. Having its origin in a world of physical products, the definition of a PE, and especially its application to modern business models such as electronic commerce that may not require a physical presence in the source jurisdiction, has recently been the subject of considerable controversy.

Treaties generally provide that arm's length prices – prices that would prevail in transactions between unrelated parties – are to be used in valuing transactions (including financial transactions) between related parties. While determining transfer prices is relatively straightforward for some homogeneous commodities that are widely traded (such as oil and wheat), it is difficult – or even conceptually impossible – for many unique intangible assets that have no market outside a given multinational corporation. The OECD has issued guidelines for the determination of transfer prices, some of which rely on formulas, but has steadfastly refused to sanction formulary apportionment (OECD 1995). Countries may disagree over the proper transfer prices, despite treaty provisions for mutual agreement procedures intended to resolve these and other conflicts in interpreting and applying treaties (for example, the residence of a taxpayer). In such cases double taxation may occur.

Preventing Fiscal Evasion

To prevent fiscal evasion, tax treaties provide for the exchange of information between tax authorities. For example, if a person deposits funds in a foreign bank, but does not report the resulting interest income, the fiscal authorities of that country could report the interest to their counterparts in the taxpayer's country of residence. In fact, exchange of information has been less useful than this description suggests. First, the fiscal authorities of the country of residence must identify the suspected tax cheat ('no fishing') and cannot require provision of information not collected in the normal course of operations or in violation of domestic law. Tax evaders can utilize legal entities whose identities are not known to the tax authorities of their country of residence to make investments.

Tax havens – low-tax jurisdictions that have bank secrecy and related laws that allow ownership of assets to be concealed – pose a particularly important threat to tax compliance. Since tax havens generally do not participate in treaty networks and resist exchange of information, it has been relatively simple to evade taxes by channelling investments to or through them. During the 1990s the OECD undertook a project on 'harmful tax competition' that included pressure on tax havens to exchange information.

Signalling

Some developing countries and countries in transition from socialism deviate from widely recognized standards for taxing business profits, for example by not allowing deductions for all business expenses or enacting tax laws that favour

domestic taxpayers over foreigners. The existence of a tax treaty provides a signal to potential investors that the signatories will play by the internationally accepted 'rules of the game', including taxation of *net* business income and non-discrimination, and assurance that their country of residence will defend them in the event of deviations from those rules. (A taxpayer may ordinarily appeal to the 'competent authority' of its country of residence to insure that both countries are abiding by the terms of the treaty.) For example, by concluding a tax treaty with Canada (its first), Mexico demonstrated readiness to join the OECD, the North American Free Trade Association and the World Trade Organization. Also, non-discrimination rules prevent source countries from levying higher taxes on non-resident investors than on local investors, in order to benefit from the availability of foreign tax credits offered by the residence country.

Multilateral Treaties

Some advocate a World Tax Organization analogous to the World Trade Organization. Such a body might foster harmonization of income tax bases, far-reaching cooperation among tax administrations, or even equalization of tax rates. Harmonization of tax bases – and especially of tax rates – is unlikely to occur soon, if ever, because, inter alia, defining the tax base and setting tax rates are important aspects of sovereignty, there is no benchmark for the ideal income tax base, and the laws of nations exhibit numerous complex differences.

Economic Issues

At first glance the two methods of preventing double taxation of business profits have quite different economic consequences. If all residence countries employed the exemption method, all income from foreign investment in a particular source country would bear only the tax of the source country and *capital import neutrality* would prevail. On the other hand, if all residence countries taxed all foreign-source income currently and allowed credits for all source-country taxes, all investment made by residents of a particular country would bear the same tax and *capital export neutrality* would prevail. There are, however, at least three reasons why the economic effects of taxing worldwide income, with credits for foreign taxes, may resemble those of an exemption system.

First, the parent's tax on most income of foreign subsidiaries is deferred until the income is distributed. (The primary exception occurs when the residence country taxes the undistributed income of certain controlled foreign corporations currently, commonly as an anti-abuse technique.) Thus systems that ostensibly tax the worldwide income of residents and systems that exempt foreign-source income produce similar results.

Second, foreign tax credits are commonly limited to the average tax rate paid on both foreign and domestic-source income. When 'excess foreign tax credits' exist, not all taxes levied by high-tax nations can be credited, again producing results similar to those of exemption systems.

Third, the tax treaties of some nations (but not the United States) allow credits for taxes that developing countries choose not to collect, because of tax incentives such as holidays and investment incentives. Such 'tax sparing' implies that tax incentives benefit investors, as under an exemption system, rather than being neutralized by higher taxes in the residence country.

See Also

- ▶ Tax havens
- ▶ Taxation of Foreign Income
- ▶ Transfer Pricing
- ▶ World Trade Organization

Bibliography

Graetz, M.J., and M.M. O'Hear. 1997. The 'original intent' of U.S. international taxation. *Duke Law Review* 46: 1021–1109.

Kosters, B. 2004. The United Nations model tax convention and its recent developments. *Asia-Pacific Tax Bulletin* 10: 4–11.

OECD (Organisation for Economic Co-operation and Development). 1995. *Transfer pricing guidelines for multinational enterprises and tax administrations*. Paris: OECD.

OECD. 2005. *Model tax convention on income and on capital*. Paris: OECD. Online. Available at http://www.oecd.org/dataoecd/52/34/1914467.pdf. Accessed 17 Feb 2007.

OECD. Online. Available at http://www.oecd.org. Accessed 17 Feb 2007.

Vann, R.J. 1998. International aspects of income tax. In *Tax law and drafting*, ed. V. Thuronyi, Vol. 2. Washington, DC: International Monetary Fund.

Taxation and Poverty

John Karl Scholz

Abstract

Low-income US households typically pay Social Security payroll taxes, state and local sales taxes, and possibly, state income taxes. Federal income taxes in the United States and the United Kingdom, among other countries, provide tax subsidies to low-income working families, particularly those with children. These 'in-work benefits' raise the incomes of poor families, modestly increase employment, and have negligible effects on hours of work. The design and effectiveness of these provisions depend on details of the tax system, such as the unit of taxation, the degree to which people file tax returns, and the ability of the tax authority to enforce tax rules.

Keywords

Earned Income Tax Credit (USA); Hours of work; Implicit tax rates; Negative income tax; Payroll taxes; Social Security in the United States; Tax compliance; Taxation and poverty; Value-added tax; Working Tax Credit (UK)

JEL Classifications

H2

Tax systems around the world can have substantial effects on the income available to families with low-skill workers. Key factors affecting the tax burdens of poor families include the set of taxes used in the economy, the specific exemptions and deductions contained in the system, and the special provisions targeting low-income households. To discuss these issues, this article focuses primarily on the experiences of the United States, but much of the discussion applies to tax systems in other developed and developing countries. For a broader treatment of taxation in developing economies, see, for example, Burgess and Stern (1993) and Gordon and Li (2005).

The primary taxes borne by low-income US households are the Social Security payroll tax, state and local sales taxes, and in some states, state income taxes. Roughly 41 per cent of US families pay more in payroll taxes than individual income taxes. If we (appropriately) assume the employer's share of payroll taxes are borne by workers, payroll taxes exceed income taxes for 71 per cent of US families. For most low-earning individuals, the net present value of Social Security benefits still exceeds the present discounted value of taxes paid (Liebman 2002), but these families are much more likely than others to be intertemporally credit constrained, so, if the payroll tax is fully borne by workers, the 14.2 per cent combined employer–employee tax results in a substantial reduction in after-tax resources available for consumption. (14.2 per cent is the sum of the employer and employee shares of payroll taxes, which equals 15.3 per cent, divided by market earnings increased by the employer's tax share – 1.0765 – with the idea that, without the payroll tax, employers would increase wages by their share of the tax.) The perceived regressivity of the Social Security payroll tax was one factor leading to the adoption of the Earned Income Tax Credit in the mid-1970s.

Sales taxes and their international cousins, value-added taxes (VAT), also raise concerns among policymakers that they impose inappropriate burdens on low-income households. Consequently, these taxes frequently exempt items such as food, clothing, and medicine that are thought to typically compose larger shares of poor families'

budgets than is the case for other families. Zero-rating (excluding) items raises a fundamental issue in taxation. Should tax systems be designed to raise the revenues necessary for the operation of government in the most efficient way possible, leaving expenditure policy to address distributional concerns, or should taxes be designed to address equity issues directly? Exempting (or zero-rating) items in a VAT or consumption tax reduces efficiency (for example, see Ballard et al. 1987). Whether policymakers deem the exemptions as being necessary depends on political considerations and the strength of other available institutions to redistribute resources to poor families.

The federal individual income tax is conspicuously absent from my list of taxes reducing the incomes of poor families. Until around 1974, the federal income tax imposed positive average and marginal tax rates on families with incomes at the US poverty line, so, along with payroll and sales taxes, income taxes (at both the federal and, in some circumstances, the state level) reduced the incomes of low-income working families. In the absence of other tax provisions targeting low-income families or individuals, the threshold at which families began to pay income taxes was determined largely by the size of the standard deduction and exemptions, and whether these provisions were indexed for inflation.

In 1974 the difference in average tax rates, combining income and payroll taxes, between a one-adult, two-child family with income at the poverty line and a two- adult, two-child family with income three times the poverty line was 9.2 percentage points, or the difference between 13.2 per cent and 22.4 per cent. By 2005 the difference was 36.9 percentage points, or the difference between −15.3 per cent and 21.6 per cent. (These calculations are made with the NBER's TAXSIM model: see Feenberg and Coutts 1993 for a discussion of TAXSIM.)

By far the most important factor affecting the tax treatment of low-income families in the United States since 1977 has been the development and expansion of tax provisions targeted to low-income taxpayers that are 'refundable' – the Treasury pays out the value of the credit regardless of whether the taxpayer otherwise has positive tax liability. The most important of these is the Earned Income Tax Credit, though in recent years a portion of the child credit has also been made refundable. Refundable credits can result in negative average tax rates for working poor families with children.

The antecedent for current tax provisions targeting low-income families is negative income tax (see Moffitt 2004, for a nice discussion). The negative income tax (NIT) was to provide a basic income guarantee that would be clawed back as earnings increase. In the mid-1970s US policymakers came close to enacting a NIT, and its labour market and family formation effects were studied extensively in a series of closely watched, widely publicized social experiments (see, Robins 1985; Cain and Wissoker 1990 for further details).

The United States implemented an Earned Income Tax Credit (EITC) in 1975. The EITC provides a subsidy to earnings up to a specific income threshold. For example, in 2004 the EITC gave a 40 per cent earnings subsidy up to 10,750 dollars to a family with two or more children. Taxpayers with earnings between 10,750 and 14,040 dollars received the maximum credit of 4300 dollars. The maximum credit for families with one child is 2604 dollars; for childless workers it is 390 dollars. The credit was reduced by 21.06 per cent of earnings between 14,040 and 34,458 dollars. Hence, there are three distinct ranges of the EITC: the subsidy, flat and phase-out ranges of the credit.

The political appeal of the EITC, and similar programmes in other countries such as the Working Tax Credit in the United Kingdom and an EITC-like earnings subsidy to be implemented in South Korea, rests on at least two factors. First, earnings subsidies like the EITC are thought to encourage work and they are sometimes justified as part of a set of policies to 'make work pay'. There is considerable evidence that this perception is accurate: the EITC has positive employment effects so, in contrast to many alternative ways of redistributing income from higher- to lower-income families, the EITC does not substantially harm labour market incentives. Second,

by adding the EITC to an existing individual income tax, implementation costs are relatively low, particularly compared with programmes that require their own bureaucracy.

The static labour supply model implies the EITC will have an unambiguous, positive incentive effect on employment. The empirical evidence is consistent with these incentive effects: the EITC has a statistically significant and large effect on labour force participation of single women with children. Grogger (2003), for example, concludes that the EITC 'may be the single most important policy measure for explaining the decrease in welfare and the rise in work and earnings among female-headed families in recent years' (2003, p. 408). For more on the EITC, see Dickert et al. (1995), Eissa and Liebman (1996), Keane and Moffitt (1998), Ellwood (2000), Meyer and Rosenbaum (2000, 2001), and Hotz et al. (2005). Eissa and Hoynes (2004) focus on the employment and hours decisions of secondary workers in married families and find small, negative effects of the credit on work. Hotz and Scholz (2003) survey EITC research.

The static labour supply model implies an ambiguous incentive effect of the EITC on hours in the phase-in range of the credit and unambiguous negative incentive effects on hours in the flat and phase-out ranges. Studies estimating the effects of the EITC on hours of work for working households find no bunching of taxpayers at the beginning and end of the phase-out range, as might be expected if the EITC significantly affects hours and taxpayers are cognizant of the discontinuities in implied marginal tax rates generated by the credit (Liebman 1997). It is not surprising that negative effects on hours for people already in the labour market are small, since the precise relationship between the EITC and hours worked is likely to be poorly understood by most taxpayers. Most EITC recipients pay a third party to prepare their tax returns, and it is difficult to infer the implicit tax rates embodied in the credit from the look-up table that accompanies the EITC instructions. This confusion is less likely to mitigate positive participation effects, since, for these to be operative, taxpayers need only to understand that there is some tax-related bonus to work. Abundant anecdotal evidence indicates that taxpayers have this understanding (see, for example, DeParle 1999).

The UK Working Tax Credit has an interesting design feature when compared with the EITC. Instead of phasing in, it imposes an hours threshold that triggers eligibility, thereby increasing the number of households receiving positive employment and hours incentives in relation to a credit on the first dollar of earnings. All households working fewer than 16 hours will see an increase in the after-tax return to work (and, since they do not receive any credit if they have fewer than 16 hours of work, there is no incentive to 'buy' more leisure). Hours limits impose a potentially significant additional administrative burden – because hours information is typically not required to implement an income tax – so their desirable labour market incentive effects must be balanced against the additional costs that arise from administering the hours requirement. Blundell and Hoynes (2004) find the EITC seems to have a larger effect on employment than the WTC predecessor, even though average EITC benefits are somewhat smaller. This may in part be because the incentive effects of in-work benefits in the United Kingdom are dulled by integrations with the rest of the tax and benefit system.

The unit of taxation in most countries around the world is the individual, not the family as is the case in the United States. Most policymakers (including those in the United Kingdom), however, believe that it is essential to target tax benefits on the basis of family income, since it is widely believed that families pool resources when making economic decisions. UK tax authorities meet this goal by having taxpayers claim eligibility by submitting a form to the tax authorities during the year, while the claim is recalculated at the end of the year based on *family* income. The UK experience shows that it is possible to have a credit with family-based eligibility in a tax system where individuals are the unit of taxation.

Less is known about the effects of the EITC on other aspects of behaviour. Dickert-Conlin and Houser (2002) and Eissa and Hoynes (2004) provide some evidence that the EITC encourages the existence of female-headed families.

Heckman et al. (2002) examine the effects of the EITC on skill formation. While they emphasize that much more needs to be done, they reach a tentative conclusion that the EITC has little impact on average skill levels in the economy. The EITC appears to reach those who are eligible – participation rates among eligible taxpayers is high (Scholz 1994). Lastly, the EITC also suffers from high rates of non-compliance (Internal Revenue Service 2002): many taxpayers who are not eligible end up claiming and receiving the credit. There is probably a trade-off between a policy with low administrative costs, like the EITC, and high rates of non-compliance.

See Also

- ▶ Anti-poverty Programmes in the United States
- ▶ Low-Income Housing Policy
- ▶ Nutrition and Public Policy in Advanced Economies
- ▶ Poverty
- ▶ Poverty Alleviation Programmes
- ▶ Welfare State

Bibliography

Ballard, C.L., J.K. Scholz, and J.B. Shoven. 1987. The value-added tax: A general equilibrium look at its efficiency and incidence. In *The effects of taxation on capital accumulation*, ed. M. Feldstein. Chicago: University of Chicago Press.

Blundell, R., and H. Hoynes. 2004. Has 'in-work' benefit reform helped the labour market? In *Seeking a premier economy: The economic effects of the British economic reforms, 1980–2000*, ed. R. Blundell, D. Card, and R. Freeman. Chicago: NBER and University of Chicago Press.

Burgess, R., and N. Stern. 1993. Taxation and development. *Journal of Economic Literature* 31: 762–830.

Cain, G.G., and D.A. Wissoker. 1990. A reanalysis of marital stability in the Seattle- Denver income-maintenance experiment. *American Journal of Sociology* 95: 1235–1269.

DeParle, J. 1999. Once a forlorn avenue, tax preparers now flourish. *New York Times*, March 21.

Dickert, S., S. Houser, and J.K. Scholz. 1995. The earned income tax credit and transfer programs: A study of labor market and program participation. In *Tax policy and the economy*, vol. 9, ed. J.M. Poterba. Cambridge, MA: NBER and the MIT Press.

Dickert-Conlin, S., and S. Houser. 2002. EITC and marriage. *National Tax Journal* 55: 25–39.

Eissa, N., and H.W. Hoynes. 2004. Taxes and the labor market participation of married couples: The earned income tax credit. *Journal of Public Economics* 88: 1931–1958.

Eissa, N., and J.B. Liebman. 1996. Labor supply response to the earned income tax credit. *Quarterly Journal of Economics* 111: 605–637.

Ellwood, D.T. 2000. The impact of the earned income tax credit and social policy reforms on work, marriage, and living arrangements. *National Tax Journal* 53: 1063–1105.

Feenberg, D.R. and E. Coutts. 1993. An introduction to the TAXSIM model. *Journal of Policy Analysis and Management* 12: 189–194. Online. http://www.nber.org/taxsim/. Accessed 14 June 2007.

Gordon, R. and W. Li. 2005. *Tax structure in developing countries: Many puzzles and a possible explanation*. Working paper No. 11267. Cambridge, MA: NBER.

Grogger, J. 2003. The effects of time limits, the EITC, and other policy changes on welfare use, work, and income among female-headed families. *Review of Economics and Statistics* 85: 394–408.

Heckman, J.J., L. Lochner, and R. Cossa. 2002. *Learning-by-doing vs. on-the-job training: Using variation induced by the EITC to distinguish between models of skill formation*. Working paper No. 9083. Cambridge, MA: NBER.

Hotz, V.J., and J.K. Scholz. 2003. The earned income tax credit. In *Means-tested transfer programs in the United States*, ed. R. Moffitt. Chicago: University of Chicago Press and NBER.

Hotz, V.J., C. Mullin, and J.K. Scholz. 2005. *Examining the effect of the earned income tax credit on the labor market participation of families on welfare*. Mimeo, UCLA and Wisconsin. Online. http://www.ssc.wisc.edu/~scholz/Research/EITC_Draft.pdf. Accessed 14 June 2007.

Internal Revenue Service. 2002. *Compliance estimates for earned income tax credit claimed on 1999 returns*. Washington, DC: Inland Revenue Service.

Keane, M., and R. Moffitt. 1998. A structural model of multiple welfare program participation and labor supply. *International Economic Review* 39: 553–589.

Liebman, J. 1997. The impact of the earned income tax credit on incentives and income distribution. *Tax Policy and the* 11: 83–119.

Liebman, J.B. 2002. Redistribution in the current U.S. social security system. In *The distributional aspects of social security and social security reform*, ed. M. Feldstein and J.B. Liebman. Chicago and London: University of Chicago Press.

Meyer, B.D., and D.T. Rosenbaum. 2000. Making single mothers work: Recent tax and welfare policy and its effects. *National Tax Journal* 53: 1027–1061.

Meyer, B.D., and D.T. Rosenbaum. 2001. Welfare, the earned income tax credit, and the labor supply of single

mothers. *Quarterly Journal of Economics* 116: 1063–1114.
Moffitt, R.A. 2004. The idea of a negative income tax: Past, present, and future. *Focus* 23: 1–4. Online. http://www.irp.wisc.edu/publications/focus/pdfs/foc232a.pdf. Accessed 14 June 2007.
Robins, P.K. 1985. A comparison of the labor supply findings from the four negative income tax experiments. *Journal of Human Resources* 20: 567–582.
Scholz, J.K. 1994. The earned income tax credit: Participation, compliance, and antipoverty effectiveness. *National Tax Journal* 47: 59–81.

Taxation of Capital

Christophe Chamley

Taxes on the income from capital have generated a large debate for two reasons. First, the contrast between the arguments for efficiency and equity seems to be particularly sharp here. Second, there exists a variety of views on the appropriate choice of an economic model and its parametric values that are relevant for policy. In this exposition, emphasis will be put on the dynamic aspects of a uniform tax on capital in general equilibrium. In most economies the tax on capital discriminates between different sectors (corporate capital, housing) and induces some static efficiency cost. This cost will be considered only very briefly in comparison with the dynamic efficiency cost.

The Impact on Capital Accumulation

A dynamic framework is essential in the analysis of the taxation of capital income, and it is useful to recall that there are three generic models of capital accumulation. The first is the so called neoclassical model with an ad hoc specification of the saving function that depends on the flow of incomes and on the interest rate. Although there is little theoretical or empirical foundation for this form, a vague justification has been found in the argument that individuals may not optimize rationally over time, or that capital markets do not operate like standard intratemporal markets. However, the main value of this specification seems to be analytical expediency. In the second type of model, individuals optimize a life-time utility function with no bequest. The third model assumes that individuals care about the welfare of their next descendants as if they would be reincarnated in these descendants with the same utility function. A recursive argument implies that individuals act as if they would live forever.

Most dynamic studies on the taxation of capital income rely on one of these models, or a variation between these types. The models have different implications for the impact of the taxation of capital income on the level of capital accumulation and output, and for the method of evaluation of the tax on capital.

In the neoclassical model the tax reduces the net flow of saving (which is equal to the growth of capital on the balanced growth path at the natural rate), and has a negative impact on capital accumulation. The magnitude of the capital reduction is of course greater when the propensity to save from disposable income has a positive elasticity with respect to the rate of return, net of tax.

In the life cycle model, the capital stock behaves like the level of water in a bathtub. On the balanced growth path it is in equilibrium between the inflow of the savings of the younger generations and the outflow of the dissavings of the older generations. The impact of the tax depends on the elasticity of this equilibrium level with respect to the net rate of return to capital. The tax induces a decrease of the level of capital, if and only if the interest elasticity is positive (Diamond and Mirrlees 1971). A weaker condition is sufficient for a decrease of capital when there is a fixed factor of production such as land because the tax on capital induces an appreciation of land that diverts savings from capital (Chamley and Wright 1986).

When the utility function is additively separable, the value of the interest elasticity of the aggregate stock of capital that is generated by the life-cycle process in the steady state depends mainly on two parameters. The first is the short-run elasticity of saving with respect to the interest rate, which is proportional to the intertemporal

elasticity of substitution of consumption. The second is the length of an individual's horizon that determines the time span during which he can accumulate capital and consume it. When the length of an individual life tends to infinity, he has more time to save up to the point where the rate of return and the rate of time preference are equal. This implies that the elasticity of supply of the stock of capital with respect to the rate of return is larger (Summers 1981).

In the limit case where individuals have infinite lives, this elasticity of supply is infinite. The impact of the capital income tax is negative and its magnitude depends on the elasticity of the demand for capital by firms. The long-term impact of the tax may be large, but a steady state analysis may be misleading since it neglects the transition period.

Dynamic Incidence

In the short-run, an increase of the tax rate on capital falls entirely on capital income. In the neoclassical model, the dynamic impact of the tax is a lower level of capital and of the wage rate in the long run, and a higher gross rate of return. In the example of an economy where all profits are saved and all wages consumed, a well-known result is that a tax on profit with transfer to workers, lowers the level of consumption of workers in the steady state. This occurs because the tax induces a shift away from the golden rule. For more general specifications, the incidence of this is shifted at least partially to labour income (Feldstein 1974).

The concept of factor incidence loses its meaning in an economy where the optimizing behaviour of agents is fully specified. In the life-cycle model for example, every individual goes through a worker and a capitalist phase. The proper evaluation method is the analysis of the welfare impact of the tax to which we now turn.

Efficiency Cost

A first step in the computation of the welfare cost of the tax on capital is to assume that the economy is composed of a large number of identical individuals who are price takers. In the dynamic context, these individuals become families with an infinite horizon. The welfare cost of the tax is defined either with respect to lump-sum taxation, or as the differential welfare cost with respect to alternate forms of distortionary taxation.

The method of analysis is to consider a small variation of the tax rate combined with a change of lump-sum taxation or of other tax rates to keep the total revenues invariant. An important assumption is that the tax rate is constant over time. The efficiency cost is given by the difference between the levels of welfare (or its income equivalent), as measured on the dynamic path with and without the tax changes, respectively. For convenience, the original position the economy is in a steady state. An essential aspect of the method is that individuals have perfect foresight and optimize rationally over time. Other forms of expectations (such as myopic expectations or other types), may be convenient for large models but they lead to strange results. For example, a tax on capital income could correct myopic expectations and improve welfare if the level of the capital stock is lower than in the steady state.

When the tax rate is small and there are no other taxes, the welfare cost is of the second order with respect to revenues, a result that is well known for any tax. Note that the impact on the steady state level of output is of the first order. This illustrates how comparisons between steady states that ignore transitional effects can lead to large errors.

The effect of an increase of the tax rate on capital income has two components. In the short run, the level of consumption increases. In the long run the levels of capital, output and consumption are lower than in the initial steady state. The relative weights of these two effects depend on the difference between the growth rate of the economy and the discount rate. Only in the special case where these two rates are almost equal, is the transition component relatively insignificant. The comparison between steady states is then a proper evaluation method.

Consider first the case where the labour supply is fixed, and assume that there is no other tax in the original position. Two structural parameters are

important for the value of the excess burden of the tax. First, there is a positive relation between the welfare cost and the intertemporal elasticity of substitution of the utility function. This effect is related to the transition path between steady states. It vanishes when the value of the growth rate tends to that of the discount rate.

Second, the welfare cost is positively related to the elasticity of substitution between capital and labour in the production function. For plausible values of this elasticity, the relation is almost linear. The smaller the elasticity, the larger is the fraction of the tax burden shifted to labour (which is a fixed factor here). When the elasticity is equal to zero, there is no distortion.

When the labour supply is elastic, the welfare cost is larger because the tax has a negative impact on the wage rate that affects the supply of labour. The marginal efficiency cost of the capital income tax is also larger when there are other taxes in the original position.

The differential welfare cost between the taxes on capital income and labour income is not always positive. A reduction of the tax rate on capital income that is maintained over time implies a lump sum transfer to the owners of the capital in place at the time of the tax reform. Therefore, a substitution of the capital income tax by the wage tax (at rates constant overtime), is not always efficient (Auerbach.et.al. 1984; Chamley 1985).

When the difference between the growth rate and the discount rate is small, the magnitude of the lump-sum transfer to the old capital is negligible with respect to the welfare cost of other taxes. In this case the differential welfare cost between the taxes on the incomes of capital and labour is positive. Its value depends on the interest elasticity of the demand for capital by firms and it is independent of the parameters of the utility function.

The measurement of the efficiency cost of the tax on capital income is more difficult when the population of individuals is heterogeneous, since it may involve implicit or explicit interpersonal comparisons of income and equity tissues. Auerbach et al. (1983, 1986), use a lump-sum redistributive authority to isolate the efficiency cost in an overlapping generation model.

Finally, the stylized model of dynamic general equilibrium is potentially a useful tool for the evaluation of a capital tax that is raised on a specific sector (such as the corporate tax), when agents optimize over time. Preliminary estimates indicate that for a production technology with constant returns to scale, the cost of the intrasectoral misallocation may be greater than the intertemporal welfare cost.

Optimal Tax Rates and Redistribution

The standard method for the determination of a programme of efficient tax rates on capital income is to choose somewhat arbitrarily, an origin of time, and to analyse from that point on the standard second-best problem where the tax on capital income is one of the fiscal instruments.

The distortion induced by the tax on capital income increases with the interval between the moment of the announcement and the date at which the tax is actually raised because the supply elasticity of savings with respect to the rate of return increases also with time. This implies the disturbing property that the policy of second-best is time inconsistent. Since the tax on capital income has a very low efficiency cost at the beginning of the policy horizon, an arbitrary limit may have to be imposed on the tax rate for some initial interval of time so that the policy is defined.

An interesting result is that for fairly general assumptions, the long-run efficient value of the tax rate on capital income is equal to zero when the fiscal instruments are the taxes on the incomes of capital and labour, respectively. The two main assumptions are that a steady state exists in the long run, and that some individuals have an infinite horizon with an asymptotic rate of time preference equal to the social rate of time preference. The latter assumptions is satisfied when the individual's utility function satisfies the axiom of Koopmans. It does not have to be separable between periods. The steady state is locally stable for additive utility functions and values of the tax

rates that are not too large (Chamley 1986). The same result holds when the wage tax is replaced by an ad valorem consumption tax.

On the transition to the steady state the value of the tax rate on capital income is in general different from zero. For an economy where the government expenditures fluctuate, the debt has been considered as a useful instrument for tax smoothing and the minimization of the efficiency cost of raising revenues (Barro 1979). It is interesting to observe that when the efficiency cost of taxation is derived explicitly from price distortions, a tax on the income of capital (with a positive or negative rate), may perform the same function (Chamley 1980). Its role is to offset the intertemporal distortions that are caused by the variations of the tax rates on consumption and labour income that occur when the government budget has to be balanced in each period.

When individuals have finite lives and an operative bequest motive *à la* Barro, the standard Ramsey rules (Diamond and Mirrlees 1981), apply for the taxation of the savings that are used in life cycle consumption. However, the taxation of intergenerational transfers is suboptimal in the second-best (i.e. when the labour tax is an alternative to the capital income tax).

The result holds under a variety of assumptions about the heterogeneity of the population (Judd 1985), and casts some doubt on the redistributive value of capital taxation in the long run. This is in contrast to other studies that use an ad hoc specification of the processes of saving and income distribution (Stiglitz 1978). The result is not valid however when there are binding restrictions on negative bequests.

In a life-cycle framework the government that maximizes a social welfare function is chosen somewhat arbitrarily, as the representative of future generations. The level of the capital generated in the process of saving and dissaving for selfish life-cycle consumption, is not in general optimal when the welfare of future generations is taken into account in an intergenerational comparison.

An important issue here is whether the government can affect the level of capital directly through its saving or dissaving. If public saving is feasible, the efficient tax rates on the incomes of capital and labour on the dynamic path tend to values in the steady state that are determined by the Ramsey rules, and they depend on the price elasticities of the supply of labour and consumption at different instants (Pestieau 1974).

When there are binding restrictions on public saving or dissaving, they can be alleviated by exploiting the differences between the timings of the taxes on labour and capital, respectively. The capital income tax is levied later in life than the labour income tax. A government that is restricted from accumulating capital, could 'entrust' the young with some capital through a labour tax, and recover it later through the capital income tax in order to place it with the next generation and so on. The opposite policy can be used when the government wants to hide the public debt from the accountants. This 'wealth carrying' function of the tax system invalidates the standard formulate for efficient taxation (Atkinson and Sandmo 1980).

Other Issues

The analysis has so far omitted the adjustment cost of investment and the international mobility of capital. The adjustment costs reduce the possibilities for intra- or intertemporal distortions, and the potential welfare gains of tax reform. The international mobility of capital has the opposite effect. One possible formulation of adjustment costs is the q-theory of investment that has been applied for the corporate tax by Summers (1981).

The issues of adjustment cost and international mobility have been integrated recently in an analytical model by Bovenberg (1986), who finds that the welfare cost of the capital income tax is significantly larger in open economies compared to closed economies, only when the degree of international capital is very high.

See Also

▶ Capital Gains and Losses

Bibliography

Atkinson, A.B.., and A. Sandmo. 1980. Welfare implications of the taxation of savings. *Economic Journal* 90: 529–549.

Auerbach, A.J., and L.J. Kotlikoff. 1987. *Dynamic fiscal policy*. Cambridge: Cambridge University Press.

Auerbach, A.J., L.J. Kotlikoff, and J. Skinner. 1983. The efficiency gains from dynamic tax reform. *International Economic Review* 24: 81–100.

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Barro, R.J. 1979. On the determination of the public debt. *Journal of Political Economy* 87: 940–971.

Bovenberg, L.A. 1986. Capital income taxation in growing open economies. *Journal of Public Economics* 31(1986): 347–376.

Chamley, C.P. 1980. Optimal intertemporal taxation and the public debt. Cowles Foundation Discussion Paper No. 554.

Chamley, C.P. 1985. Efficient tax reform in a dynamic model of general equilibrium. *Quarterly Journal of Economics* 100: 335–336.

Chamley, C.P. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54: 607–622.

Chamley, C.P., and B.D. Wright. 1986. *Fiscal incidence in an overlapping generation model with a fixed factor*. Mimeo: Hoover Institution.

Diamond, P.A. 1965. National debt in a neoclassical growth model. *American Economic Review* 55: 1125–1150.

Diamond, P.A. 1970. Incidence of an interest income tax. *Journal of Economic Theory* 2: 211–224.

Diamond, P.A., and J.A. Mirrlees. 1971. Optimal taxation and public production. II: Tax rules. *American Economic Review* 61: 261–278.

Feldstein, M. 1974. Incidence of a capital income tax in a growing economy with variable savings rates. *Review of Economic Studies* 41: 505–513.

Judd, K.L. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 28: 59–83.

Pestieau, P. 1974. Optimal taxation and discount rate for public investment. *Journal of Public Economics* 3: 217–235.

Stiglitz, J.E. 1978. Notes on estate taxes, redistribution and the concept of balanced growth path incidence. *Journal of Political Economy* 86: S137–S150.

Summers, L. 1981a. Capital taxation and capital accumulation in a life cycle growth model. *American Economic Review* 71: 533–544.

Summers, L. 1981b. Taxation and corporate investment: A *q*-theory approach. *Brooking Papers on Economic Activity* 1: 67–127.

Taxation of Corporate Profits

Alan J. Auerbach

Abstract

Corporate profits taxes account for a relatively small share of revenues in leading industrial countries but represent a potentially important source of economic distortion. The incidence of corporate taxes has traditionally been assigned to owners of capital, but more recent theories have suggested that many other groups, from shareholders to owners of other domestic factors of production, may share the burden, and that the burden itself may be overstated. Although commonly described as taxes on income, corporate profits taxes may have quite different bases, making the economic effects potentially quite different from those of a tax on corporate source income.

Keywords

Capital accumulation; Capital gains and losses; Capital gains taxation; Capital mobility; Computable general equilibrium model; Deadweight loss; Debt finance; Depreciation; Dividend taxation; Double taxation; Efficiency effect of taxation; Equity finance; Excise taxes; Harberger, A. C.; Imperfect competition; Insurance; Investment risk; Portfolio investment; Progressive and regressive taxation; Proportional taxation; Tax avoidance; Tax competition; Tax distortions; Tax incidence; Taxation of capital income; Taxation of corporate profits; Vintage capital

JEL Classifications

H2

This is a revised version of the article by Peter Mieszkowski in the first edition of the dictionary.

Industrial countries commonly levy a tax on the earnings of corporations.

Corporate income taxes account for a relatively small share of revenues in these countries. As of 2000, the share of total government revenues accounted for by the corporate tax (OECD 2002, Table 13) reached a high among the G-7 countries in Japan, at 14 per cent, and a low in Germany, at five per cent. In a number of these countries, this share had dropped substantially from the 1960s. For example, the US share stood at 16 per cent of revenues in 1965, almost double its 2002 value of nine per cent. The decline in importance of corporate tax revenues has been attributed to a variety of overlapping factors, including tax policy, shifts in business activities out of corporate form, more aggressive tax avoidance behaviour, financial innovation, and increasing tax competition and capital mobility among jurisdictions. In sorting through these potential explanations, economists continue to develop models of corporate tax incidence and efficiency effects and to consider the rationale for a separate tax on corporate income.

The US corporate tax dates at least to a corporate profits excise tax in 1909, four years before a constitutional amendment cleared the way for the 1913 introduction of a general income tax that included a corporate profits tax. In the years since, two key measures have been the relative taxation of distributed and retained earnings and the relative taxation of corporate and non-corporate capital income. The original US income tax imposed a 'normal' one per cent tax rate on both retained and distributed earnings but also imposed a graduated individual income surtax of up to six per cent on a tax base that excluded retained corporate earnings, leaving retained earnings subject to a much lower overall tax rate than distributed earnings and leaving corporate income as a whole facing a lower tax rate than income from non-corporate sources, which could not escape the surtax. Through the years, however, although the tax rate on retained earnings generally remained below that on distributed earnings, increases in the corporate level tax caused both rates to rise relative to tax rates on non-corporate income (Bank 2006). By the post-Second World War era, the corporate income tax was viewed as imposing an extra tax burden on activities within the corporate sector, even with the favourable treatment of retained earnings. This set the stage for the extremely influential analysis by Harberger (1962 and 1966, respectively) of the incidence and efficiency effects of the corporate income tax.

Harberger's Contributions

Dividing the US economy into two sectors according to whether production was predominantly carried out by corporate or non-corporate businesses, Harberger (1962) characterized the corporate tax as an additional tax levied on capital income originating in the corporate sector, layered on top of the individual income tax collected on capital income from both sectors. He then estimated incidence through the changes in factor prices and product prices that would result from a small increase in the corporate tax. Harberger's main conclusion was that, under reasonable assumptions regarding the two sectors' production elasticities of substitution and consumers' elasticity of substitution between the two sectors' products, the corporate income tax was borne fully by owners of capital, economy-wide. This finding has two important elements. First, capital bears the entire tax; it is not shifted to labour or consumers. Second, it is all capital, not just corporate capital, that bears the tax.

Harberger's second contribution (1966) is his estimates of the inefficiency resulting from the higher tax imposed on corporate capital. As a result of the corporate tax, he saw the social (before-tax) rate of return as higher in the corporate sector. Real national income would increase if the tax distortion were eliminated and capital were reallocated to equalize rates of return across sectors. Harberger estimated the deadweight loss of the tax differential between corporate and non-corporate investment to be about seven per cent of the taxes collected on corporate earnings.

Harberger's analyses spawned a vast literature that extended and challenged his initial results.

The simplicity of Harberger's technique – comparative static analysis of small changes in a two-sector model – proved not to be a major source of concern given that similar findings resulted from analysis using a multi-sector computable general equilibrium model (Shoven 1976). But Harberger also relied on several other simplifying assumptions including: (*a*) free mobility of factors across sectors; (*b*) that the corporate tax can be viewed as an add-on tax on capital income originating in the corporate sector; (*c*) no risk; (*d*) competitive markets and constant returns to scale; (*e*) a closed economy; and (*f*) fixed economy-wide factor supplies. These and other assumptions have been examined in the literature.

Dynamics

It is reasonable to think of the shifts predicted by the Harberger model as occurring over time with after-tax returns to capital slowly equalizing in the two sectors and with corporate assets initially dropping in value to reflect the gap in after-tax returns, consistent with the q-theory of investment envisioned by Tobin (1969) and developed by Hayashi (1982), Summers (1981) and others. Thus, a corporate income tax with gradual adjustment to Harberger's long-run outcome would be borne partially by current owners of corporate capital, through an initial drop in asset values, and partially by future investors in corporate and non-corporate capital, through lower rates of return. The inefficiency of the corporate tax would be changed by gradual adjustment, with weaker responsiveness to tax changes translating into smaller present-value deadweight losses.

Investment Provisions

As modelled by Harberger, the base of the corporate income tax equals income from all corporate capital. In particular, income from capital goods of different vintages is taxed at the same rate. In reality, capital goods of different ages receive different treatment, even though they are subject to the same statutory corporate tax rate, because of differences in depreciation provisions and other incentives provided to new investment. With accelerated depreciation and investment incentives, new assets are more attractive than old ones of the same productivity. Put another way, the effective tax rate on new investment may be lower than the effective tax rate on existing capital. The differential should be capitalized into the value of existing assets.

Calculations in Auerbach (1983) found that the value ratio of old to new corporate fixed capital in the United States should have been around 0.8 in the early 1980s, with a reduction in both investment incentives and the statutory corporate rate reducing capitalization substantially by the next decade (Auerbach 1996). Thus, there is a second component of the corporate tax borne by corporate asset holders rather than by all capital, and potentially lower deadweight loss as well to the extent of the tax wedge being shifted from new capital to existing assets. Auerbach (1983) found the marginal effective corporate tax rate in the United States to be well below the average effective corporate rate.

Corporate Financial Policy and Taxation

With corporations having the option to issue debt, the interest payments which are deductible at the corporate level, and to retain earnings, thereby trading off current dividends for capital gains on which taxes may be lower and can be deferred, how much 'double taxation' does corporate capital actually face? In the extreme, if corporations finance all their investment by borrowing, there is no corporate tax imposed on investment; indeed, corporate tax liability is reduced, because nominal interest payments – a portion of which simply compensates lenders for a loss in purchasing power – are tax deductible.

Based on these attributes, Stiglitz (1973) concluded that firms should pay no dividends, retain all their earnings, and meet any additional financing needs by issuing debt. While this theory explains why some equity would exist, it fails to

explain why so much equity exists. Here, the theory of Miller (1977) comes in. Miller focused on the heterogeneity of individual investors, arguing that, under a progressive tax system, there may be some investors in high enough tax brackets that the extra taxation at the corporate level is more than offset by the preferential individual tax treatment of equity income. Under Miller's theory, investors with a tax preference for equity would hold equity, those with a tax preference for debt would hold debt, and corporations would be indifferent between the two, issuing enough of the two securities to satisfy the demands of investors.

Even with investor heterogeneity, is it plausible that a significant share of investors will have a tax preference for equity? A very low effective equity tax rate would be required, and this seems inconsistent with the fact that a substantial share of equity earnings comes to investors as fully taxed dividends. One potential explanation is that many countries provide some form of dividend relief, either through reduced taxation at the corporate level (through a lower rate of corporate tax on distributed earnings or a deduction for dividends paid) or at the investor level (through a lower rate of tax on dividends received or a shareholder imputation credit for taxes paid at the corporate level). An alternative explanation comes from the 'new view' of dividend taxation (Auerbach 1979; Bradford 1981; King 1977), under which the effective rate of individual tax on equity income may be the capital gains rate, adjusted for deferral – a very low rate – even if dividends are distributed, when retained earnings are the source of equity finance, as they are for most large corporations. This view stands in stark contrast to the 'traditional' view, under which the effective individual tax rate on equity earnings is a weighted average of the tax rates on dividends and capital gains, the weights reflecting the shares of corporate earnings distributed and retained.

The new view, which also attempts to explain why mature firms pay dividends, is based on the intuition that existing equity funds are 'trapped' within the firm, unable to get out easily without being subject to a tax rate on dividends. As a consequence, the dividend tax rate (or, more precisely, the excess of the dividend tax rate over the effective individual capital gains tax rate on retained earnings) will be capitalized into share values, having no effect on the incentive to distribute earnings or on the effective tax rate on equity-financed investment. Through the years, different empirical strategies have been used to test the relative validity of the traditional and new views of the impact of equity taxation. One approach, based on the q-theory investment model, appeared to provide strong support for the traditional view when based on UK data (Poterba and Summers 1985) but equally strong support for the new view when based on US data (Desai and Goolsbee 2004). Other approaches focusing on rates of return (Auerbach 1984) and the source of investment funds (Auerbach and Hassett 2003) have suggested the presence of firm heterogeneity, with the new view more relevant for 'mature' firms with ample internal funds.

Regardless of whether there are investors with a sufficiently low tax rate on equity, another serious challenge to the Miller model is that investors clearly do not specialize. Tax-exempt institutional entities invest substantially in equity, and higher-income individuals hold at least some corporate bonds. As discussed by Auerbach and King (1983), the Miller model breaks down when assets are risky and investors must balance the objectives of diversification and tax minimization. Tax preferences will influence portfolios – those in higher brackets will still gravitate more towards assets, like equity, with more favourable individual tax treatment. This modification of the model implies that the incidence conclusions based on the simple Miller model are overly strong; while high-bracket investors suffer more from an increase in the corporate tax, even tax-exempt investors will also bear some of the burden.

Risk

Since the work of Domar and Musgrave (1944), economists have noted that taxes on capital income provide insurance as well as imposing burdens. As has been established in the literature,

a proportional tax system that provides a full loss offset (that is, the same tax rate applies whether income is positive or negative) imposes a burden on investors only to the extent that the safe return is taxed. This result, combined with the empirical observation that the real, safe rate of return is very close to zero, led Gordon (1985) to suggest that the corporate income tax imposes few economic distortions and has little incidence, although it collects tax revenue on average (that is, in expected value).

Gordon's conclusion was challenged by Bulow and Summers (1984), who argued that, while the government shares in the income risk of a corporate investment, much of the investment risk is associated with fluctuations in the price of capital goods, and the corporate tax base excludes accruing capital gains and losses on fixed capital. Likewise, the corporate tax's risk-sharing is reduced by imperfect loss offsets, which also raise the effective tax rate. Thus, even if a pure, symmetric tax on accruing corporate income caused few distortions, this is unlikely to be true of actual corporate tax systems that consistently raise substantial revenues from corporations.

Imperfect Competition

Harberger's conclusions on incidence were challenged by the econometric results of Krzyzaniak and Musgrave (1963), who, on the basis of time-series analyses of American manufacturing, reached the startling conclusion that *after-tax* profits rose in the short run in response to increases in the corporate tax rate. The Krzyzaniak–Musgrave contribution was criticized by a number of writers, but their empirical finding is possible under certain forms of imperfect competition. The study's methodology does not allow one to identify the nature of corporate responses, but corporations behaving in an oligopolistic manner need not maximize joint profits, and therefore might increase before-tax profits, and possibly even after-tax profits, by reducing output and hence increasing prices in response to an increase in the corporate tax.

The International Economy

Unlike in the purely domestic context, there is a distinction between where income is earned and where its owner resides, and the concept of residence, itself, is applied not only to individuals but also to corporations. Countries may seek to tax corporate income on a source basis, a residence basis, or some combination of the two, and most countries follow this last approach, taxing at least some income at source at the corporate level, even if the corporation is owned abroad, and taxing at least some portfolio income of domestic residents on holdings of foreign assets.

As in the analysis underlying the Miller model, an equilibrium with individuals possessing different relative tax preferences for different assets leads to specialization of the highest-bracket investors in the most tax-favoured assets (Gordon 1986), but the number of possible allocations of assets among investors is increased by the fact that individuals may hold foreign assets in many countries and in a variety of ways (for example, portfolio investment versus direct investment), and corporations (and, to a lesser extent, individuals) can change the location not only of their investments but also of their tax residence.

Among the key results in the international tax context is that a corporate income tax will be partially shifted to non-capital domestic factors of production, the more so the smaller is the taxing jurisdiction (Kotlikoff and Summers 1987). Also, with many taxing jurisdictions, the possibility of 'tax competition' arises, with governments setting their corporate tax rates strategically in response to the tax polices of other countries. In this context, a 'race to the bottom' is a possible, though by no means a certain, outcome. But the reductions in corporate tax rates in recent decades provide some evidence for a strengthening of tax competition (Devereux et al. 2002).

The Long Run

One of the Harberger model's most important omissions is the impact of corporate income

taxes on capital accumulation. We would expect an increase in the effective tax rate on new saving and investment to reduce capital accumulation. The resulting decline in the capital–labour ratio would increase before-tax returns to capital and lead to a fall in wages, thus partially shifting the tax burden from capital to labour, with much the same effect as capital flight in the open economy. This analysis would apply to the corporate tax as well, but only to the extent that the corporate income tax represents a tax on new saving and investment.

See Also

- ▶ Dividend Policy
- ▶ Tax Competition
- ▶ Tax Incidence

Bibliography

Auerbach, A.J. 1979. Share valuation and corporate equity policy. *Journal of Public Economics* 11: 291–305.

Auerbach, A.J. 1983. Corporate taxation in the United States. *Brookings Papers on Economic Activity* 1983(2): 451–505.

Auerbach, A.J. 1984. Taxes, firm financial policy and the cost of capital: An empirical analysis. *Journal of Public Economics* 23: 27–57.

Auerbach, A.J. 1996. Capital allocation, efficiency and growth. In *Economic effects of fundamental tax reform*, ed. H. Aaron and W. Gale. Washington, DC: Brookings Institution.

Auerbach, A.J., and K.A. Hassett. 2003. On the marginal source of investment funds. *Journal of Public Economics* 87: 205–232.

Auerbach, A.J., and M.A. King. 1983. Taxation, portfolio choice and debt–equity ratios: A general equilibrium model. *Quarterly Journal of Economics* 98: 588–609.

Bank, S.A. 2006. A capital lock-in theory of the corporate income tax. *Georgetown Law Journal* 94: 889–947.

Bradford, D.F. 1981. The incidence and allocation effects of a tax on corporate distributions. *Journal of Public Economics* 15: 1–22.

Bulow, J.I., and L.H. Summers. 1984. The taxation of risky assets. *Journal of Political Economy* 92: 20–39.

Desai, M.A., and A.D. Goolsbee. 2004. Investment, overhang, and tax policy. *Brookings Papers on Economic Activity* 2004(2): 285–355.

Devereux, M.P., R. Griffith, and A. Klemm. 2002. Corporate income tax reforms and international tax competition. *Economic Policy* 17: 450–495.

Domar, E.D., and R.A. Musgrave. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58: 388–422.

Gordon, R.H. 1985. Taxation of corporate capital income: Tax revenues versus tax distortions. *Quarterly Journal of Economics* 100: 1–27.

Gordon, R.H. 1986. Taxation of investment and savings in a world economy. *American Economic Review* 76: 1086–1102.

Harberger, A.C. 1962. The incidence of the corporation income tax. *Journal of Political Economy* 70: 215–240.

Harberger, A.C. 1966. Efficiency effects of taxes on income from capital. In *Effects of corporation income tax*, ed. M. Krzyzaniak. Detroit: Wayne State University Press.

Hayashi, F. 1982. Tobin's marginal and average q: A neoclassical interpretation. *Econometrica* 50: 213–224.

King, M.A. 1977. *Public policy and the corporation*. London: Chapman & Hall.

Kotlikoff, L.J., and L.H. Summers. 1987. Tax incidence. In *Handbook of public economics*, vol. 2, ed. A. Auerbach and M. Feldstein. Amsterdam: North–Holland.

Krzyzaniak, M., and R.A. Musgrave. 1963. *The shifting of the corporation tax*. Baltimore: Johns Hopkins Press.

Miller, M. 1977. Debt and taxes. *Journal of Finance* 32: 261–275.

OECD (Organization for Economic Cooperation and Development). 2002. *Revenue statistics of OECD member countries 1965–2001*. Paris: OECD.

Poterba, J.M., and L.H. Summers. 1985. The economic effects of dividend taxation. In *Recent advances in corporate finance*, ed. E.I. Altman and M.G. Subrahmanyan. Homewood: Richard D. Irwin.

Shoven, J.B. 1976. The incidence and efficiency effects of taxes on income from capital. *Journal of Political Economy* 84: 1261–1283.

Stiglitz, J.E. 1973. Taxation, corporate financial policy and the cost of capital. *Journal of Public Economics* 2: 1–34.

Summers, L.H. 1981. Taxation and investment: A *q* theory approach. *Brookings Papers on Economic Activity* 1981(1): 67–127.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1: 15–29.

Taxation of Foreign Income

James R. Hines

Abstract

Taxation of foreign income entails the taxation by one country of income that its residents earn

in another country. While most countries exempt active foreign business income from taxation, several large capital exporters subject foreign income to taxation but permit taxpayers to claim credits for taxes paid to foreign governments. There is extensive empirical evidence that the taxation of foreign income influences the magnitude of foreign investment and the tax avoidance activities of investors. Neutral taxation of foreign income entails considerations not only of the volume and location of investment, but also the effects of taxation on capital ownership.

Keywords

Capital export neutrality; Capital import neutrality; Capital ownership neutrality; Double taxation; Foreign investment; National neutrality; Tax credits; Tax harmonization; Taxable income; Taxation of corporate profits; Taxation of foreign income

JEL Classifications

H2

Taxation of foreign income entails the taxation by one country of income that its residents earn in another country.

Most countries subject some types of foreign income to taxation. Since this income is also typically taxed by the foreign countries in which it is earned, there is considerable scope for ruinous double taxation. For example, in the 1970s the corporate tax rate in the United States was 48 per cent while the corporate tax rate in Germany was 56 per cent; without some attenuation of double taxation, the combined tax rate of 104 per cent would probably have discouraged (profitable) American corporate investment in Germany.

International practice since the dawn of taxation is that countries tax income earned within their borders, whereas countries in which taxpayers are resident grant tax relief for foreign income in order to reduce or eliminate double taxation. There are two primary alternative methods by which residence countries grant relief, the first being to exempt foreign income from taxation, and the second being to permit residents to claim credits for taxes paid to foreign governments. Many countries combine these systems, exempting active foreign business income from taxation while subjecting foreign personal income to taxation but permitting individuals to claim credits for income taxes paid to foreign jurisdictions.

The nature of international commerce is such that most foreign income is earned by businesses rather than individuals. Many countries largely exempt active foreign business income from taxation, though a number of major capital exporting nations, including the United States, the United Kingdom and Japan, tax foreign income while granting credits for taxes paid to foreign governments. With such a system of taxing foreign income, and a home-country corporate tax rate of 35 per cent, a corporation that earns 100 in a foreign country that imposes ten per cent tax rate pays taxes of 10 to the foreign government and 25 to its home government, since its home-country corporate tax liability of 35 is reduced to 25 by the foreign tax credit of ten. Since foreign tax credits are intended to alleviate international double taxation, credits are limited to home-country taxes due on foreign income; taxpayers are not permitted to use taxes paid to foreign governments to reduce home-country tax liability on domestic income. In addition, countries that tax active foreign income permit taxpayers to defer home-country taxation of certain business profits earned and reinvested abroad; that income is taxed only when repatriated to the country of residence.

Effects of Taxing Foreign Income

The taxation of foreign income and the tax laws of other countries have the potential to influence a wide range of corporate and individual behaviour, including, most directly, the location and scope of international business activity. Studies of behavioural responses to international tax rules find that multinational firms invest less in high-tax countries than they do in otherwise-similar low-tax countries. This is most evident from the disproportionate shares of financial and real

investment in tax haven countries (Hines 2005), but also appears in cross-sectional econometric estimates of the determinants of foreign investment. Controlling for income levels and other observable characteristics of host countries, foreign direct investment levels are negatively associated with local corporate tax rates, the implied elasticity of investment with respect to the tax rate generally lying close to −0.6 in data covering the 1980s (Hines and Rice 1994), and increasing in magnitude to −1 or greater in evidence data since the 1990s (Altshuler et al. 2001). High rates of local taxes other than corporate income taxes are likewise associated with reduced levels of foreign investment (Desai et al. 2004a).

There is extensive evidence that firms arrange financial flows and intrafirm sales to reallocate taxable income from high-tax countries to low-tax countries. This reallocation is commonly accomplished by concentrating corporate borrowing, and therefore interest deductions, in high-tax countries (Desai et al. 2004b), and by adjusting prices paid for intrafirm financial transactions and sales of goods and services to minimize income reported in high-tax countries (Clausing 2003). As a consequence, multinational firms report significantly higher profit rates in low-tax countries than in high-tax countries (Desai et al. 2003), and the ability to reallocate taxable income only increases the attractiveness of investing in low-tax countries.

Taxation of foreign income, together with provision of foreign tax credits, dampens incentives to earn income in low-tax countries, since lower foreign tax payments reduce available foreign tax credits and thereby create greater home-country tax obligations. Foreign investment in the United States is consistent with these incentives, in that investors from countries that exempt foreign income from taxation concentrate their investments more heavily in low-tax states than do investors from countries that tax foreign income (Hines 1996). The taxation of foreign income restricts the attractiveness of investment in low-tax countries to situations either in which ample foreign tax credits are available, or in which investors can profitably defer home-country taxation. In practice, American firms are much more likely to reinvest foreign profits earned in low-tax locations, since immediately returning these profits to the United States would produce significant tax obligations (Desai et al. 2001).

The impact of home-country taxation is illustrated by the practice of granting 'tax sparing' credits for investments in certain developing countries, thereby permitting taxpayers to claim credits for normal rates of foreign taxes, whether or not these have been actually paid. Evidence indicates that Japanese investors are much more likely to receive local tax concessions in countries with which Japan has 'tax sparing' agreements than they are elsewhere, and that Japanese investment is concentrated in these countries as a result (Hines 2001). Finally, the taxation of foreign income has even encouraged some individuals and multinational firms to expatriate, effectively changing their places of tax residence to avoid home-country taxation of lightly taxed foreign income (Desai and Hines 2002).

Neutral Taxation of Foreign Income

International tax rate differences may encourage inefficient allocation of economic activity; consequently, considerable effort has been devoted to understanding the properties of tax systems that create neutral incentives.

Capital export neutrality (CEN) is the doctrine that an investor's income should be taxed at the same total rate regardless of the location in which it is earned. If a home-country tax system satisfies CEN, then a firm seeking to maximize after-tax returns has an incentive to locate investments in a way that maximizes pre-tax returns. This allocation of investment promotes global economic efficiency under certain circumstances. The CEN concept is frequently invoked as a normative justification for taxation of foreign income with provision of foreign tax credits (Richman 1963), though in practice, countries limit foreign tax credits and commonly defer taxation of unrepatriated active business income.

The same logic implies that governments acting on their own, without regard to world welfare, should want to tax the foreign incomes of their

resident companies while permitting only deductions for foreign taxes paid. Such taxation satisfies what is known as national neutrality, discouraging foreign investment by imposing a form of double taxation, but doing so in the interest of the home country, which disregards the value of tax revenue collected by foreign governments. From the standpoint of the home country, foreign taxes are simply costs of doing business abroad, and therefore warrant the same treatment as other costs. This line of thinking suggests that countries fail to advance their own interests in permitting taxpayers to claim foreign credits, or worse, in exempting foreign income from taxation.

A third neutrality principle is capital import neutrality (CIN), the doctrine that the return to capital should be taxed at the same total rate regardless of the residence of the investor. Pure source-based taxation at rates that differ between locations can be consistent with CIN, since different investors are taxed at identical rates on the same income. In order for such a system to satisfy CIN, however, it is also necessary that individual income tax rates be harmonized, since CIN requires that the combined tax burden on saving and investment in each location should not differ between investors. While CEN is commonly thought to characterize tax systems that promote efficient production, CIN is thought to characterize tax systems that promote efficient saving (Horst 1980).

The importance of ownership for productivity, and the reality that much foreign investment consists of acquisitions of existing assets by new owners, has prompted analysis of the features of tax systems that do not distort ownership of capital. Capital ownership neutrality (CON) is satisfied if every country taxes foreign income similarly, thereby avoiding tax-based ownership clienteles (Desai and Hines 2003). From the standpoint of capital ownership, a country fails to advance world welfare by adopting a tax system that promotes CEN, if most capital exporters exempt foreign income from taxation.

The same circumstances that make CON desirable from the standpoint of world welfare also imply that countries acting on their own have incentives to exempt foreign income from taxation, regardless of what other countries do. The reason is that additional outbound foreign investment does not reduce domestic activity, since reduced home-country investment by domestic firms is offset by greater investment by foreign firms. Home-country welfare rises with the productivity of domestic factors, and is maximized by ownership patterns produced by exempting foreign income from taxation. Tax systems that exempt foreign income from taxation are therefore said to satisfy national ownership neutrality. Hence it is possible to understand why so many countries exempt active foreign business income from taxation, and it follows that, if every country did so, capital ownership would be allocated efficiently, to the benefit of global productivity.

See Also

- ▶ Neutral Taxation
- ▶ Tax Competition
- ▶ Tax Havens
- ▶ Taxation of Corporate Profits

Bibliography

Altshuler, R., H. Grubert, and T.S. Newlon. 2001. Has U.S. investment abroad become more sensitive to tax rates? In *International taxation and multinational activity*, ed. J.R. Hines Jr.. Chicago: University of Chicago Press.

Clausing, K.A. 2003. Tax-motivated transfer pricing and US intrafirm trade prices. *Journal of Public Economics* 87: 2207–2223.

Desai, M.A., C.F. Foley, and J.R. Hines Jr. 2001. Repatriation taxes and dividend distortions. *National Tax Journal* 54: 829–851.

Desai, M.A., C.F. Foley, and J.R. Hines Jr. 2003. Chains of ownership, tax competition, and the location decisions of multinational firms. In *Foreign direct investment in the real and financial sector of industrial countries*, ed. H. Herrmann and R. Lipsey. Berlin: Springer.

Desai, M.A., C.F. Foley, and J.R. Hines Jr. 2004a. Foreign direct investment in a world of multiple taxes. *Journal of Public Economics* 88: 2727–2744.

Desai, M.A., C.F. Foley, and J.R. Hines Jr. 2004b. A multinational perspective on capital structure choice and internal capital markets. *Journal of Finance* 59: 2451–2487.

Desai, M.A., and J.R. Hines Jr. 2002. Expectations and expatriations: Tracing the causes and consequences of corporate inversions. *National Tax Journal* 55: 409–440.
Desai, M.A., and J.R. Hines Jr. 2003. Evaluating international tax reform. *National Tax Journal* 56: 487–502.
Hines Jr., J.R. 1996. Altered states: Taxes and the location of foreign direct investment in America. *American Economic Review* 86: 1076–1094.
Hines Jr., J.R. 2001. Tax sparing and direct investment in developing countries. In *International taxation and multinational activity*, ed. J.R. Hines Jr. Chicago: University of Chicago Press.
Hines Jr., J.R. 2005. Do tax havens flourish? In *Tax policy and the economy*, vol. 19, ed. J.M. Poterba. Cambridge, MA: MIT Press.
Hines Jr., J.R., and E.M. Rice. 1994. Fiscal paradise: Foreign tax havens and American business. *Quarterly Journal of Economics* 109: 149–182.
Horst, T. 1980. A note on the optimal taxation of international investment income. *Quarterly Journal of Economics* 94: 793–798.
Richman, P.B. 1963. *Taxation of foreign investment income: An economic analysis*. Baltimore: Johns Hopkins University Press.

Taxation of Income

Alan J. Auerbach

Abstract

Income taxes are the single most important source of revenue for most countries, although there is an active debate about the relative attractiveness of alternatives such as broad-based consumption taxes. In practice, the income tax base deviates from a comprehensive income measure in several important respects, by excluding non-market activities, limiting refunds for losses, and including capital gains on realization rather than on accrual. Each deviation introduces additional distortions of taxpayer behaviour. Defining the family unit for purposes of income taxation remains a complex issue, as does the optimal degree of tax progressivity.

Keywords

Capital gains taxation; Consumption taxation; Depreciation; Direct taxation; Double taxation; Expenditure taxation; Haig–Simons measure of income; Indexation of the income tax; Indirect taxation; Inflation; Interpersonal utility comparisons; Marginal tax rates; Negative taxation; Progressive and regressive taxation; Realized vs unrealized capital gains; Tax distortions; Taxation of income

JEL Classifications

H2

Despite its current prominence as a government revenue source, the income tax's first appearance is relatively recent in the historical evolution of government finance.

Evidence of a serious national income tax is difficult to find before the end of the 18th century, when William Pitt achieved the passage in Great Britain of the Act of 1799, which imposed a comprehensive income tax, complete with exemptions and abatements for dependents, on all residents of Great Britain. Introduced to maintain the British government's solvency during the Napoleonic Wars, the income tax was dispatched once the French had been. Seligman (1911, p. 113) quotes a contemporaneous source as stating that the repeal of the tax by Parliament 'was declared amidst the greatest cheering and the loudest exultation ever witnessed within the halls of the English Senate'. It was only decades later that the income tax reappeared in Britain.

This pattern of introduction during wartime, followed by repeal and eventual, permanent reinstitution is found in the experience of other countries, as well. In the United States, for example, the first income tax was introduced during the civil war in 1862, being abandoned in 1872. It reappeared in 1894 in similar form, but was almost immediately declared unconstitutional by the Supreme Court, which found it to be a 'direct' tax not apportioned among the states according to population. The 16th amendment to

the constitution was required before the income tax could be imposed again in the United States, in 1913.

Over time, the income tax has grown in importance so that it now represents the single most important revenue source in most developed countries. In 2000 (according to OECD 2002, Table 9), for example, taxes on income and profits among the G-7 countries (the United States, Japan, Germany, France, United Kingdom, Italy, and Canada) accounted for between roughly one quarter and one half of all revenues, with the United States relying the most on the income tax (51 per cent) and France the least (25 per cent).

Being a direct tax on individuals, rather than an indirect tax on transactions, the income tax requires a more developed government infrastructure than other revenue sources. This distinction also provides the key to understanding both why the income tax was seen as a fairer way to raise revenue and why it was so vehemently opposed. Through assessment of individuals, the income tax was better suited to the achievement of a progressive, broad-based tax structure than the agglomeration of indirect taxes and duties that preceded it. At the same time, this focus on individuals instead of transactions brought with it the perception of a challenge to individual liberty, both because of the exposure to the government of the individual's economic behaviour and the ability of government to levy arbitrarily high taxes on small groups of taxpayers (see, for example, Blum and Kalven 1953).

The Measurement of Income

Dating back almost to the introduction of the income tax itself is the question of how income should be measured. What has come to be called the 'Haig–Simons' measure of income is now generally accepted as the appropriate base for an income tax (Haig 1921; Simons 1938). As expressed by Simons (1938, p. 50), 'Personal income may be defined as the algebraic sum of (1) the market value of rights exercised in consumption and (2) the change in the value of the store of property rights between the beginning and end of the period in question.' One may justify the Haig–Simons approach on grounds of both fairness and efficiency, the former because it treats individuals with different sources of income uniformly and the latter because it does not distort decisions of how to devote resources to the generation of income.

Yet actual income taxes vary from this definition, for reasons of both administration and politics.

Market Versus Non-market Activities

A range of activities generates imputed income that, by the Haig–Simons definition, should be included in the income tax base. Some sources of imputed income, such as imputed rent from owner-occupied housing, have been seriously considered for inclusion in the tax base. At the other extreme are various home production activities such as cleaning and home repair.

The inability and unwillingness of governments to tax income from non-market activities introduces a distortion of taxpayer choices, for it encourages the taxpayer to substitute non-market for market activities. Such substitution may be entirely legal, as in the purchase or cleaning of one's own home, or illegal, as in the establishment of professional cooperatives wherein members provide services to other members 'for free'.

Realizations Versus Accruals

The Haig–Simons measure does not distinguish between realized and unrealized increases in wealth. Yet most income tax systems include capital gains in the tax base only when they are realized, if at all. This outcome is traceable in part to the difficulty of measuring unrealized gains, although this can hardly be a problem for the vast wealth held in marketable securities. A related difficulty, often ascribed to farmers and the owners of small businesses, involves illiquid assets which would have to be sold below their going-concern values were their owners subject to taxes on associated accrued gains. Finally, property rights to capital assets may be sufficiently vague or disputed that it is difficult even to attribute ownership until gains have actually been realized. A variety of proposals to tax such

accrued gains retrospectively upon realization (for example, Vickrey 1947; Auerbach 1991) have attracted little more than academic attention.

This favourable treatment of capital gains facilitates the accumulation and transmission of wealth. It has therefore been viewed as mitigating the progressivity of the income tax system, leading some (for example, Kaldor 1955) to favour an individual expenditure tax on grounds of equity. At the same time, others have argued that favourable capital gains treatment serves to encourage risk-taking, which is otherwise discriminated against by the income tax system (see the discussion below).

Nominal Versus Real Income

Income has been viewed as an appropriate measure of the individual's ability to pay, but this ability is generally viewed in real rather than money terms, an individual being no better off with twice the income at twice the price level. Price-level indexation of the income tax requires two types of corrections: to the rate structure and to the base itself. The first is required because of the progressivity of marginal tax rates, the second because capital income is measured incorrectly in the presence of inflation.

When the income tax structure has marginal tax rates that rise with nominal income, increases in the price level raise both the marginal and the average tax burden on a given real income level. Correction for inflation involves indexing tax brackets to the price level, a practice implemented only in 1985 in the United States, after a period of relatively high inflation.

Capital income is generally mismeasured in an inflationary environment because changes in the real value of capital goods that are due to inflation are generally treated incorrectly by the tax system. This has led to four problems identified by the literature (see, for example, Aaron 1976); the understatement of costs of goods sold from inventories, the understatement of the depreciation expenses associated with the use of durable capital goods, the overstatement of income received from bonds and other nominal commitments, and the overstatement of realized capital gains. In all cases, the problem arises from a failure to apply the Haig–Simons approach to changes in the value of capital assets. There have been few attempts to ameliorate these distortions in practice. In fact, some economists have suggested that maintaining these apparently gratuitous distortions serves a positive purpose, namely, to weaken government's appetite to pursue inflationary policies (Fischer and Summers 1989).

Losses

There is no logical reason why the income tax base for an individual cannot be negative, but treating losses as gains are treated would call for a negative tax payment, that is, a government refund. This outcome is rarely observed in practice, in part perhaps because restricting the use of losses imposes some limit on the extent to which taxpayers can fraudulently under-report income. Instead, individuals with currently negative tax bases are permitted to average the current base retrospectively or prospectively with the aim of achieving a net positive number. In the case of forward averaging (called 'carrying forward') this still amounts to the penalty of having to wait for the refund until future taxes are due without any interest to compensate for this deferral.

A closely related outcome occurs under a progressive tax structure when income is positive in every year, but fluctuates from year to year. Since marginal rates are higher in good years than bad (a milder version of what occurs when negative tax bases face a tax rate of zero), taxpayers face a higher tax in present value than if they received the same present value of income, but in a smooth stream over time. As with the treatment of losses, tax systems typically provide some imperfect form of averaging of incomes over several contiguous years to lessen this problem.

This treatment of losses and risky incomes has been viewed as discouraging the taking of risks (for example, Domar and Musgrave 1944), and has been one of the more valid reasons for favouring the preferential treatment normally accorded capital gains. At the same time, there is no general presumption that income taxation, in itself, discourages the taking of risks, since it reduces not only the returns to risky investments but also the risks. In fact, the reduction in risk may increase

private risk-taking (Domar and Musgrave 1944; Tobin 1958), but this outcome may be reversed if government cannot reduce the risk of its resulting revenue stream (Gordon 1985).

Defining the Unit of Taxation

In addition to these problems of income measurement, ambiguities have arisen concerning the delineation of the taxpaying unit. Questions have concerned how broadly to define a unit at a given date, and over what time interval to measure the income accruing to that unit.

Tax Treatment of the Family

Tax systems vary in their treatment of related individuals. The method of treatment of family members affects the tax burden on a family because of the progressivity of the rate structure. Two individuals will generally be assessed a different tax bill if considered separately than if taxed jointly as a couple, regardless of how the rate structure is adjusted, since the total tax bill under separate taxation will depend on the distribution of taxable income between the individuals while under joint taxation it will not.

Even if families can be identified and grouped for purposes of taxation, the problem remains in deciding how to vary the tax schedule with family size. For this, one must have a measure of how to normalize income by family size to obtain a measure of the family's ability to pay. Such questions have been addressed but not often applied to the design of tax schedules.

Finally, how the family is grouped also matters if the tax-free transfer of resources through gifts and bequests is not allowed. The strict Haig–Simons approach would include gifts and bequests in the tax base of the recipient, but such transfers might not be observed if occurring within the unit of taxation.

Tax Treatment over Time

In Simons's own description of the appropriate measurement of income, the element of time plays a crucial role, since accretions to wealth must be defined over some interval. Indeed, the difference between a tax on income and a tax on expenditures amounts to a different choice of time interval over which to measure income for purposes of taxation. If, instead of annual income, we assessed taxes on lifetime income, then individuals would pay taxes in excess of lifetime consumption only to the extent that they accumulated resources over their lifetimes. Taking the further step of assessing families rather than individuals, we might ignore even the lifetime resource accumulation that would go to finance the consumption of subsequent generations.

This point has not been missed by advocates of the expenditure tax, who argue that, from a lifetime perspective, individual expenditures are a better measure of ability to pay than annual income. Under the annual income tax, individuals who consume their resources later in life face a heavier lifetime burden, paying taxes when income is initially earned and again when interest on the saved capital is received in later years. This outcome led to the charge that the annual income tax imposes unfair 'double taxation' of savings, an argument made by early proponents of a consumption tax (for example, Fisher 1939; Kaldor 1955). More recent arguments in favour of consumption taxation have focused on considerations of economic efficiency. The central point of this literature has been that, notwithstanding the government's inability to avoid all tax-induced distortions, a system including capital income taxation, which imposes greater and greater distortions on consumption as the time horizon lengthens, cannot be optimal (Chamley 1986; Judd 1985).

On the Optimal Progressivity of the Income Tax

As soon as the income tax became ensconced as a revenue source, economists began to consider how progressive it should be. Early researchers in the utilitarian tradition focused on how rapidly the tax burden should rise with income so as to exact an equal sacrifice from each individual given the particular utility function with which people were assumed to be endowed. The answer also depended on whether one was seeking equal

absolute sacrifice, equal proportional sacrifice, or equal marginal sacrifice, all measured in units of the interpersonally comparable individual utilities (Musgrave 1959, pp. 99–105).

Perhaps the most disquieting result from this line of investigation was that the achievement of equal marginal sacrifices required the equalization of incomes across individuals, if utility function were the same (Edgeworth 1897). A missing element that would have altered this finding was the distortionary impact of taxes on economic behaviour. The high marginal tax rates needed to approach equality of after-tax incomes (100 per cent in the extreme case of complete equality) would undoubtedly become self-defeating, in that the after-tax incomes of all individuals would begin to fall as tax revenues ceased increasing with increases in marginal tax rates. Such an outcome would be inconsistent with any evaluation of social welfare that had Pareto efficiency as a necessary condition for an optimum.

There followed, eventually, a line of research that sought to reconcile the utilitarian aim of equal marginal sacrifice with the disincentive effects of marginal taxation. The seminal paper here is that of Mirrlees (1971), who found optimal marginal rates to be relatively low. Subsequent research has also shown that marginal tax rates should eventually approach zero at the highest incomes, a result that is not only dependent on various assumptions but also less politically controversial than one might first expect once it is recognized that it applies to marginal, not average, tax rates, and possibly only at very high incomes.

Further research on the distortions of the income tax has focused on its efficiency relative to other tax bases, such as labour income and consumption expenditures. A key result here is that of Atkinson and Stiglitz (1976), who showed that, under a relatively plausible restriction on preferences, government could not improve on a progressive tax on labour income using differential consumption taxes. If one interprets different commodities as consumption at different dates, then the implication is that a tax on labour income, or equivalently a tax on lifetime consumption, is the optimal progressive tax. Many complications limit the direct application of this result in the determination of actual tax policy, but the lesson has been helpful in what remains an active area of research. One notable complication is that tax policy evolves over time. This evolution makes comparisons of tax systems different from evaluations of a transition from one tax system to another, in which the treatment of assets accumulated under the previous tax system must be taken into account in performing welfare analysis (Auerbach and Kotlikoff 1987).

See Also

- ▶ Capital Gains Taxation
- ▶ Consumption Taxation
- ▶ Optimal Taxation
- ▶ Public Finance

Bibliography

Aaron, H., ed. 1976. *Inflation and the income tax*. Washington, DC: Brookings Institution.

Atkinson, A., and J. Stiglitz. 1976. The design of tax structure: Direct versus indirect taxation. *Journal of Public Economics* 6: 55–75.

Auerbach, A. 1991. Retrospective capital gains taxation. *American Economic Review* 81: 167–178.

Auerbach, A., and L. Kotlikoff. 1987. *Dynamic fiscal policy*. Cambridge: Cambridge University Press.

Blum, W., and H. Kalven. 1953. *The uneasy case for progressive taxation*. Chicago: University of Chicago Press.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54: 607–622.

Domar, E., and R. Musgrave. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58: 388–422.

Edgeworth, F. 1897. The pure theory of taxation. *Economic Journal* 7: 550–571.

Fischer, S., and L. Summers. 1989. Should governments learn to live with inflation? *American Economic Review* 79: 382–388.

Fisher, I. 1939. The double taxation of savings. *American Economic Review* 29: 16–33.

Gordon, R. 1985. Taxation of corporate capital income: Tax revenues versus tax distortions. *Quarterly Journal of Economics* 100: 1–27.

Haig, R. 1921. The concept of income: Economic and legal aspects. In *The federal income tax*, ed. R. Haig. New York: Columbia University Press.

Judd, K. 1985. Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics* 28: 59–83.

Kaldor, N. 1955. *An expenditure tax*. London: Allen & Unwin.
Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.
Musgrave, R. 1959. *The theory of public finance*. New York: McGraw-Hill.
OECD (Organization for Economic Cooperation and Development). 2002. *Revenue statistics of OECD member countries 1965–2001*. Paris: OECD.
Seligman, E.R.A. 1911. *The income tax*. New York: Macmillan.
Simons, H. 1938. *Personal income taxation*. Chicago: University of Chicago Press.
Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.
Vickrey, W. 1947. *Agenda for progressive taxation*. New York: Ronald Press.

Taxation of the Family

James Alm

Abstract

In administering an individual income tax, a country must decide what constitutes an 'individual'. This choice has traditionally been seen as one between making either the family or the individual the 'unit of taxation'. The choice between the family and individual as the unit of taxation in the income tax – indeed in any tax or transfer programme – is not clear-cut, and involves difficult trade-offs between competing and worthwhile goals. This article examines some of the issues that countries face in choosing the unit of taxation, or what is often referred to as 'taxing the family'.

Keywords

Community property laws; Equity; Horizontal equity; Income splitting; Labour supply; Marriage and divorce; Marriage tax; Progressive and regressive taxation; Tax compliance costs; Taxation of income; Taxation of the family; Vertical equity; Women's work and wages

JEL Classifications

H2

Nearly all countries around the world impose an individual income tax. In administering this tax, each country must decide what constitutes an 'individual'; that is, each country must choose the 'unit of taxation'. This choice has traditionally been seen as one between the family and the individual. In the former case, the incomes of all members of a family are aggregated, and the income tax (with all of its relevant provisions) is then imposed on total family income. In the latter case, each individual is taxed only on his or her own individual income, even if he or she is a member of a family unit in which other members have taxable income.

The choice between the family and individual as the unit of taxation is not clear-cut, and involves difficult trade-offs between competing and worthwhile goals. With the dramatic increase in recent years of different household 'types' – cohabiting but not legally married couples, extended families, same-sex couples, unrelated individuals living together – these issues have become even more complicated. The presence of numerous other tax and transfer programmes whose magnitude is determined by family status complicates these issues still more.

This article examines some of the issues that countries face in choosing the unit of taxation, or what is often referred to as 'taxing the family'.

Some Goals and Principles in Taxing the Family

Countries have a variety of goals in choosing the structure of the individual income tax. Such a tax is usually viewed as balancing the various desirable attributes of taxation: taxes must be raised *(adequacy)* in a way that treats individuals fairly (*equity*) and in a way that minimizes interference in economic decisions (*efficiency*). See Boskin and Sheshinski (1983) and Apps and Rees (1999) for an analysis of the optimal taxation of the family.

Defining equity is quite difficult. One notion of equity requires that taxpayers with greater income pay greater amounts of taxes. It is generally felt that a progressive rate structure is best able to achieve vertical equity, sometimes referred to as the *progressivity* goal of taxation.

Another notion requires that taxpayers who are equal in all relevant respects pay equal amounts of taxes. The difficulty here lies in defining 'equals'. Equals can be defined in terms of married couples with equal income, but also as any 'household' with equal income. If a married couple is seen as the relevant household type for defining equals, then achieving the goal of *horizontal equity across households* requires that married couples with equal incomes pay equal taxes. However, if a household is defined more broadly, then achieving this goal requires that any households with equal income pay the same amount of taxes, requiring the additional and separate goal of *equal payments by singles and couples.* This goal can easily be broadened to apply to all household types.

Still another goal is *marriage neutrality*, which requires that a couple's combined income tax liability remain unchanged with marriage. However, there is substantial evidence that many couples pay more in taxes as a married couple than their combined taxes as single individuals; this is often referred to as a 'marriage tax' or 'marriage penalty'. There is also evidence that marriage can reduce tax liabilities, in which case the reduction in taxes is called a 'marriage subsidy' or 'marriage bonus'. As discussed later, it is well documented that the US individual income tax is not marriage neutral, and exhibits a large and variable marriage penalty – and marriage bonus – whose magnitude has changed over time. There is also evidence that the income tax is not marriage-neutral in many other countries.

In general, any income tax can create a marriage penalty or subsidy if two conditions are satisfied: the tax is based on household income, and the tax imposes different marginal tax rates at different levels of income (Steuerle 1999). Many tax and transfer programmes meet these conditions, so that that marriage non-neutrality exists throughout the fiscal system in almost all countries. For the United States, the General Accounting Office (1996) identified 59 provisions in the individual income tax code that contribute to a marriage penalty or subsidy, and over 1,000 federal laws in which marital status is a factor in the determination of taxes or transfers, ranging from income tax and welfare provisions to programmes involving veterans' payments, immigrant benefits, and other social insurance programmes.

It is now well-known that no individual income tax can achieve the simultaneous goals of *horizontal equity across families*, *equal payments by singles and couples, progressivity,* and *marriage neutrality.* Choosing the features of the individual income tax therefore requires that countries must face trade-offs in their pursuit of worthwhile goals. Countries have made very different choices in this regard.

Worldwide Practice in Taxing the Family

In the case of the United States, the tax treatment of the family has varied over time (Bittker 1975; Berliant and Rothstein 2003). The basic unit of income taxation was initially the individual. However, after the Second World War a growing number of states instituted community property laws, which allowed married couples to divide their income equally and file separate tax returns, thereby giving a significant tax advantage to couples living in community property states. In response, the Revenue Act of 1948 changed the unit of taxation from the individual to the family with the adoption of 'income splitting' for married couples, in which all couples were allowed to aggregate and to divide in half their income for federal tax purposes. Due to the progressive nature of the individual income tax, a couple's joint tax liability fell with marriage (for example, a marriage bonus).

This marriage bonus grew over the next two decades. Public pressure to remedy this disparity let to the adoption of the Tax Reform Act of 1969, which established a new, separate tax schedule for single individuals that insured that single persons would incur a maximum tax liability of 120 per cent of a married couple with equal income.

However, a side effect of this reform was the creation, for the first time, of a marriage tax or penalty for many married couples, especially for couples with similar earnings. Since then, various tax and demographic changes have markedly affected the potential for a marriage penalty or subsidy, as well as the magnitude of each. In the longer run, the size of the marriage penalty will be heavily influenced by the alternative minimum tax.

Other countries have made very different choices in taxing the family (Alm and Melnik 2005). In most OECD countries (as of 2002 tax laws), the individual is the unit of taxation, and joint filing for couples is not permitted. Joint filing is required in only seven countries (Belgium, France, Greece, Luxembourg, Portugal, Switzerland, and the United States), while six countries allow couples to select the filing status (Germany, Iceland, Ireland, Norway, Poland, and Spain). A total of 17 OECD countries use only the individual as the unit of taxation, and, as noted, another six countries in which the taxpayer can choose between single or joint taxation. Not every country that permits or requires joint filing allows joint assessment (for example, income splitting between the spouses). For instance, Greece, Norway, and Spain all have provisions for joint filing, but income splitting does not apply, which means that joint filing is not meaningfully different from single filing, except when joint filing allows a couple to use different personal exemptions. Income splitting is present in some form in only nine of the 32 OECD countries. In most countries that allow income splitting, the income of the spouses is simply aggregated, so that the tax system does not differentiate between households with equal combined incomes based on how the income is distributed within the couple. However, there are exceptions to this as well. For example, Belgium allows only limited income splitting, which applies only to those couples in which there is a significant differential between the spouses' incomes.

Income splitting in the presence of a progressive tax rate structure creates a tax benefit to couples when spouses earn different incomes, as evidenced quite clearly by the US experience. Furthermore, the tax benefit is a function of the difference in those incomes and the marginal tax rate structure. For instance, Luxembourg has narrowly defined marginal tax rate brackets, and a relatively small differential can translate into a significant tax saving for a married couple as opposed to two single taxpayers with similar incomes, as long as neither spouse falls into the highest income bracket. In contrast, in Iceland and Ireland a much larger difference in incomes may have no impact on tax liability due to the marginal tax rate structure.

Many OECD countries also have special tax provisions that apply to single-earner couples, in an attempt to provide some form of tax relief to these couples. The most common provision is some form of credit, deduction, allowance, or rebate (for example, Australia, Austria, Canada, the Czech Republic, Denmark, Iceland, Italy, Japan, the Republic of Korea, and the Slovak Republic). Another popular provision is income splitting, used in Belgium, France, Germany, Ireland, Luxembourg, Poland, Portugal, and the United States.

In general, the dominant practice of individual income taxation in OECD countries is to choose the individual rather than the family as the unit of taxation, and thereby to tax individuals on their own income even if they are married. As a result, the individual income tax is largely marriage neutral in these countries. This practice of taxing the individual is one that has tended to emerge since the mid- 1970s in these countries. Even so, there remains much diversity in how OECD countries choose to tax the family.

Some Effects of Taxing the Family

Defining the taxable unit as the family rather than the individual is controversial. The principal arguments revolve around equity issues. However, there are also efficiency issues, as well as revenue effects.

The basic economic model of marriage indicates that income taxes may affect the gains to marriage via two paths. First, differential income tax treatment of married couples may alter the

total taxes paid by the couple relative to taxes paid as single individuals. If total taxes paid increase (decrease) with marriage, *ceteris paribus*, then the gains to marriage unambiguously fall (rise). Second, marriage may change the marginal tax rate faced by the couple relative to that faced as singles. A higher marginal tax rate with marriage increases the tax liability of the couple and so lowers the benefits of marriage; however, a higher marginal tax rate also lowers the after-tax wage rates of the individuals, thereby reducing the opportunity cost of household production work and increasing the gains from marriage. The tax system therefore both creates incentives for those marriages that involve traditional gender roles and discourages those marriages with two wage earners.

Much recent empirical work has attempted to disentangle the effects of income taxes on marital decisions, with most of this work focusing on the United States. See Whittington and Alm (2003) for a summary of many of these studies. This work has tended to find that the marriage tax has a small but statistically significant impact on marriage and divorce probabilities. The income tax may also affect the timing of the marriage decision, and several studies have found that couples in the United States and in Canada, England and Wales have timed their marriages to avoid one year of the tax penalty. There is some evidence that income tax influences the likelihood that individuals live together as a legally married versus as a cohabiting couple. Even so, in most instances the magnitude of the tax impact is relatively small. For example, Alm and Whittington (1999) find that at mean values a ten per cent rise in the marriage penalty leads to a 2.3 per cent reduction in the possibility of first marriage, and Alm and Whittington (1997) estimate that doubling the tax penalty increases the probability that a couple delays its marriage to the next tax year by one per cent but that the tax penalty subsidy has no impact on the timing of divorce.

Marriage penalties and subsidies also affect the labour supply decisions (both participation and hours) of married individuals. Consider the secondary earner of the family. With the family as the unit of taxation, the secondary earner in a couple will be taxed at the marginal tax rate faced by the family on its combined income, and this tax rate is likely to be much higher than the tax rate that the individual would face if single. Taxing the family thereby discourages both labour force participation and hours worked of the secondary earner. Recent estimates of labour supply elasticities indicate that female labour force participation is especially responsive to marginal tax rates, particularly for women in high-income couples. See Kniesner and Ziliak (2008) for a comprehensive survey.

In sum, this recent research clearly demonstrates that the marriage penalty/ subsidy distorts some individual decisions, even if these effects are not always large.

A number of other studies have attempted to calculate the actual penalties or subsidies in the US individual income tax. Although the precise estimates differ, their broad outlines are generally the same. For example, Feenberg and Rosen (1995) use individual tax returns to compute the US marriage penalty for 1994. They find that 51 per cent of married couples paid an average marriage penalty of about $1,200, 38 per cent received an average marriage subsidy of $1,400, and 11 per cent were unaffected. Couples more likely to incur a marriage penalty were two-earner families (especially with similar incomes), families with children, families with higher family income, and older families; single-earner couples generally received a marriage bonus. However, there was much dispersion in the size of the penalty/ bonus across households. In total, income taxes were $6 billion higher than otherwise. Studies by Alm and Whittington (1996), the Congressional Budget Office (1997), and Bull et al. (1998) give comparable results.

The marriage penalty/bonus affects families at different levels of income very differently. For example, Alm and Whittington (1996) find that families in the highest family income quintile had an absolute tax penalty that was unambiguously larger than that borne by low- or middle-income couples but that, as a percent of income, the penalties and subsidies were much larger for low-income families. The Congressional Budget Office (1997) estimates that the average tax

penalty for a family with less than $20,000 in adjusted gross income (AGI) was 7.6 per cent, compared with 1.6 per cent for families with AGI greater than $50,000. However, the level of family income does not unambiguously determine the magnitude of the marriage penalty/bonus. Rather, it is the distribution of income between husband and wife that largely determines the magnitude and the sign of the change in taxes with marriage.

Overall, these differential tax treatments of individuals and families introduce large, variable, and capricious inequities due to unequal treatment of taxpayers based solely on their marital status:

- between married couples with one earner (who get a bonus relative to what they would pay as singles) and those with two earners (who pay a penalty), even if each type of couple pays the same taxes as a married couple;
- between married couples and cohabiting couples – individuals may choose to live together as unmarried cohabitors because of the income tax savings that unmarried status gives them;
- between married couples and single households (for example, the so-called 'singles tax'); and
- between married couples and extended households (for example, non-marital cohabitation in same-sex households, in households with related individuals, in households with unrelated individuals).

As Steuerle (1999) has noted, the marriage tax is almost like a voluntary tax, imposed on those who have decided voluntarily to marry.

Conclusions

What do we want an individual income tax to achieve? There is today an enormous, and increasing, diversity of family structures in many countries. In the 1950s, the 'traditional family' was typically a single-earner household with a stay-at-home spouse. Now, many individuals choose to live alone, two-earner families are the norm, non-marital cohabitation among opposite and same-sex couples is common, extended families are increasing in numbers, and there are widespread instances of unrelated individuals living together. These newer types of households are, by many definitions, 'families'. However, they are treated very differently, and often much less favourably, than the traditional households once envisioned as the norm by the tax codes in many countries.

If concern with the marriage penalty/bonus is an overriding issue, it is certainly possible to move the individual income tax toward *marriage neutrality.* One obvious method here is to make the individual the unit of taxation. However, it is also possible to move closer to neutrality even while retaining the family as the unit of taxation, by such piecemeal reform options as increasing the standard deduction for married couples, expanding the tax brackets facing married couples, establishing a secondary earner deduction, expanding the phase-out range of transfer programmes, flattening overall rate structures, or allowing optional individual filing. More fundamental reform options include eliminating progressivity (for example, a flat rate tax) or even eliminating the individual income tax and replacing it with a national sales tax or a value-added tax.

However, these reform options would only reduce the marriage penalties (and bonuses) in the individual income tax. There would still be marriage penalties and bonuses throughout other parts of the tax and transfer system.

Moreover, suppose the individual is made the unit of taxation, as many OECD countries have chosen to do. This choice is not without its own efficiency, equity, and adequacy problems. An important justification for the use of the family as the unit of taxation is the notion that families with equal family income should pay equal taxes. There is no question that making the individual the unit of taxation would violate this goal of *horizontal equity across families,* as well as *equal payments by singles and couples.* There are also significant administrative and compliance issues from individual taxation. How are itemized deductions split between partners? How is

unearned (or capital) income split between partners? Who claims the tax benefits from children? How do the tax enforcement agencies verify the legitimacy of these declarations? What are the compliance costs of individual filing? Many other such issues naturally arise, and the ways in which these issues are resolved vary greatly across countries.

It may well be that the importance of the traditional family unit still justifies its favourable tax treatment. This is clearly the avenue that the United States has chosen, and it seems unlikely that this choice will change anytime soon. However, it may also be time to recognize that a diverse society can no longer treat one family structure so differently from the others. Making the individual the unit of taxation would eliminate the marriage tax/subsidy (and the singles tax), and would also reestablish the principle of horizontal equity, broadly defined to apply to individuals and not to families or to couples. Many countries have in fact chosen to make the individual the unit of taxation.

There are no easy choices here, and it is inevitable that the goals of taxation are often conflicting. Taxing the family requires facing these difficult trade-offs directly.

See Also

- ▶ Family Decision Making
- ▶ Family Economics
- ▶ Horizontal and Vertical Equity
- ▶ Marriage and Divorce
- ▶ Marriage Markets
- ▶ Taxation of Income

Bibliography

Alm, J., and M.I. Melnik. 2005. Taxing the 'family' in the individual income tax: An international perspective. *Public Finance and Management* 5: 67–109.

Alm, J., and L.A. Whittington. 1996. The rise and fall and rise... of the marriage tax. *National Tax Journal* 49: 571–589.

———. 1997. Income taxes and the timing of marital decisions. *Journal of Public Economics* 64: 219–240.

———. 1999. For love or money? The impact of income taxes on marriage. *Economica* 66: 297–316.

Apps, P.F., and R. Rees. 1999. Individual versus joint taxation in models with household production. *Journal of Political Economy* 107: 393–403.

Berliant, M., and P. Rothstein. 2003. Possibility, impossibility, and history in the origins of the marriage tax. *National Tax Journal* 56: 303–317.

Bittker, B.I. 1975. Federal income taxation and the family. *Stanford Law Review* 27: 1388–1463.

Boskin, M.J., and E. Sheshinski. 1983. Optimal tax treatment of the family. *Journal of Public Economics* 20: 281–297.

Bull, N., Holtzblatt, J., Nunns, J.R., and Rebelein, R. 1998. Assessing marriage penalties and bonuses. Working Paper. Washington, DC: Office of Tax Analysis, U.S. Department of the Treasury.

Congressional Budget Office. 1997. *For better or for worse: Marriage and the federal income tax*. Washington, DC: Congress of the United States.

Feenberg, D.R., and H.S. Rosen. 1995. Recent developments in the marriage tax. *National Tax Journal* 48: 91–101.

Kleven, H.J., Kreiner, C.T., and Saez, E. 2006. The optimal income taxation of couples. Working Paper No. 12685. Washington, DC: NBER.

Kniesner, T.J., and J.P. Ziliak. 2008. Evidence of tax-induced individual behavioral responses. In *Fundamental tax reform: Issues, choices, and implications*, ed. J.W. Diamond and G.R. Zodrow. Cambridge, MA: MIT Press.

Steuerle, C.E. 1999. Valuing marital commitment: The radical restructuring of our tax and transfer systems. *The Responsive Community* 9: 35–45.

US General Accounting Office. 1996. *Income tax treatment of married and single individuals*, GAO/GGD-96-175. Washington, DC: US General Accounting Office.

Whittington, L.A., and J. Alm. 2003. The effects of public policy on marital status in the United States. In *Marriage and the economy: Theory and evidence from advanced industrial societies*, ed. S. Grossbard-Shechtman. New York: Cambridge University Press.

Taxation of Wealth

Alan J. Auerbach

Abstract

One of the oldest forms of taxation, wealth taxes may take the form of annual taxes based on property or net worth, or assessments

collected at less regular intervals, on estates or inheritances or in the form of emergency capital levies. Wealth taxes are still prominent but have become less important than income taxes as a source of revenue. Although wealth taxes are related in structure to taxes on capital income, their economic effects depend on their form. In addition to explicit taxes on wealth, governments impose implicit capital levies through changes in tax policy.

Keywords

Capital levies; Dynamic inconsistency; Estate taxes; George, H.; Implicit wealth taxes; Inheritance taxes; Land tax; Local wealth taxes; Net worth taxation; Precautionary savings; Property tax; Single tax; Taxation and risk-taking; Taxation of capital income; Taxation of wealth; Tiebout hypothesis

JEL Classifications

H2

Wealth taxation is one of the oldest methods of government revenue collection, having been used at least since the time of the ancient Greeks.

According to Seligman (1895, p. 34), Athens levied a general property tax not only on land and houses, but also on slaves, cattle, furniture, and money. The succeeding centuries have seen many types of wealth taxation tried and others proposed, with perhaps no other form of taxation being the subject of such heated debate.

Wealth taxes are in some sense taxes on capital income. All assets have value because of the returns they generate, though the returns need not be in explicit form (as with business assets) but may be implicit (as with owner-occupied housing or gold). Taxing wealth on an annual basis may therefore be viewed as equivalent to taxing its return at a rate sufficient to produce the same tax revenue. However, a number of factors make wealth taxation and capital income taxation different in practice.

First, governments often levy taxes on very particular forms of wealth (such as land), while income taxation has more typically been broad-based. Second, taxes on income have usually been confined to income explicitly realized. This is, for example, the nearly universal approach to capital gains taxation. Wealth taxes, such as property taxes on owner-occupied real estate, have not followed the same principle. Thus, it has been possible for taxpayers with very little liquidity but substantial wealth to be burdened with taxes well in excess of their *explicit* income. Moreover, wealth is generally even more unevenly distributed among the population than realized capital income. Finally, while it is natural to expect that income taxes would not exceed 100 per cent of income, there is no comparable upper bound on wealth taxes short of the entire stock of wealth. The perception that wealth taxes threaten the rights of individual citizens through the possibility of discriminatory or unfair taking of assets is undoubtedly a factor in the controversy that has often surrounded wealth taxation.

As of 2000, wealth taxes accounted for five per cent of tax revenues for the average OECD country, although the share was ten per cent or more in several countries, including the United States, the United Kingdom, Canada and Japan (OECD 2002, Table 23).

Types of Wealth Taxation

In discussing the history and economic effects of wealth taxation, it is useful to distinguish the primary forms of wealth taxation that have been used. There are four that may be considered important.

Property Taxation

This is the oldest form of wealth taxation, dating from antiquity. It is characterized by a tax at regular intervals (for example, yearly) on particular forms of private wealth, most commonly land, but also other forms of property. Originating in an era when land ownership was a much better measure of one's ability to pay than is currently so, property taxes have gradually been replaced and supplemented by other forms of taxation. In the United States, for example, where property taxes are still relatively important, the fraction of

government revenue coming from property taxation fell to ten percent or so in the years after the Second World War, compared with around 40 per cent during the first three decades of the 20th century (Carter et al. 2006, Figure Ea-c).

Estate and Inheritance Taxation

Taxes on bequests and inheritances first appeared long after general property taxes. They differ from property taxes in that they are assessed only once, at death, and typically apply to most assets bequeathed.

Estate taxes are typically quite complicated and not particularly successful either at revenue collection or wealth redistribution. Some have referred to the estate tax as a 'voluntary tax' because there are so many tax-planning devices available to reduce or eliminate the tax burden imposed on transfers to wealth.

Net Worth Taxation

As discussed above, net worth taxation, which applies to assets net of liabilities, is very similar in effect to a tax on capital income, differing primarily in its coverage of assets which do not generate substantial current realized income.

Whereas property taxes and taxes on estates or inheritances are quite common, only a handful of developed countries currently supplement their revenue collections with broad-based annual taxes on net worth. At the end of the 20th century, only two OECD countries, Iceland and Spain, reported non-negligible revenues from taxes on net worth, although even for these two countries such taxes still accounted for less than one half of one per cent of total revenues (OECD 2002).

Capital Levies

In wartime, countries need vast resources over short periods of time. The preferred method of raising such funds has been the issuance of national debt, but several countries resorted in the 20th century to the capital levy, a 'one time only' tax on existing wealth, more burdensome than annual net worth taxes but ostensibly temporary. Such levies were imposed in the First World War by Germany, Czechoslovakia, Austria and Hungary, and prior to the Second World War by Italy and Hungary (Hicks et al. 1941). Naturally, fully unanticipated, onetime wealth taxes are non-distortionary if not particularly fair, but the appearance of one country more than once in the above list indicates the difficulty of using capital levies unexpectedly, a problem popularly known as 'dynamic inconsistency'.

The Economic Effects of Wealth Taxation

One may start an analysis of the effects of wealth taxation by noting its similarity to the taxation of capital income, with its coincident discouragement of saving. In practice, however, different forms of wealth taxation may have different or additional effects because of their design. For example, the capital levy may be less distortionary to the extent that it is unanticipated.

The 'Single Tax'

In *Progress and Poverty* (1882), Henry George argued for the use of a tax on the rent of unimproved land as the chief source of government revenues. Though some contemporary authors disagreed, it is fairly clear that George's tax, in hitting the return to a productive factor in extremely inelastic supply, would have imposed minimal distortions of economic behaviour. Of course, the 'single tax' movement that followed for many years after was invested with a fervour based on much more than the desire for Pareto efficiency.

Taxation and Risk-Taking

Many authors (for example, Domar and Musgrave 1944; Stiglitz 1969) have noted the potentially different impact on risk-taking occurring under capital income and wealth taxes because of the failure of the latter to vary with the asset returns actually realized. But this distinction is more apparent than real once investors have taken the opportunity to scale their asset holdings up or down in response to taxation. As Tobin (1958) showed, a tax on risky capital income has no effect on individual welfare and private risk-taking when the safe rate of return is zero. It may further be shown (Gordon 1985) that, when social

risks are efficiently traded, neither does social risk-taking change when such a tax is levied.

These results are quite easily extended to the case when the safe rate of return is positive and the tax is on capital income in excess of this safe return on assets. Thus, if we view a tax on all capital income as one on the safe return to capital and one on the excess returns that compensate for risk, only the former component has economic impact. Hence, a capital income tax may be seen as equivalent to a tax levied on asset values multiplied by a fixed, safe rate of return, in which case it is obviously equivalent to a tax on wealth (Kaplow 1994). Note that this equivalence requires efficient risk-bearing. Otherwise, capital income taxation may be preferred because it allows the government to pool risks that individual investors have not been able to pool privately.

The Tiebout Hypothesis

In the United States, property taxes are used primarily to pay for local public services, and are the dominant source of finance for such services. This connection means that, if communities differ in their property tax burdens, one cannot ignore the implied differences in public services if it is necessary to live in a community to partake of its public services, such as education, trash removal, and police and fire protection. In the absence of such a connection, one would expect property tax differentials to be reflected in land prices. However, with government using the taxes to pay for desired public outputs, one might expect a different result.

This notion was formalized by Tiebout (1956), who sketched a theory in which different communities levy different taxes and provide different bundles of public services, the result being a Pareto-optimal allocation with individuals choosing their place of residence according to the bundle of local public goods and services desired. Aside from the difference between such an entrance fee and the actual property tax, there are many other issues that arise in considering the validity of the Tiebout model (Mieszkowski and Zodrow 1989). Nevertheless, local wealth taxes, with competing jurisdictions, are clearly different in their impact from national wealth taxes.

Bequests and the Estate Tax

If annual wealth taxes discourage saving in a way similar to annual capital income taxes, one might presume the same result for estate taxes, with the large anticipated one-time burden having an important impact on lifetime saving. However, this presupposes an economic model of bequests that is by no means well accepted, that is, that they are the manifestation of a desire to leave resources to heirs, and that they are influenced by the 'price' of an after-tax dollar in the heir's hands. At least one factor leading to bequests is the absence of efficient annuities markets, inducing the need for elderly individuals to engage in precautionary saving to provide for unanticipated health expenses or a longer than predicted lifetime. Such saving would not be influenced at all by taxes imposed after death. Hence, one might view estate taxes as being potentially less distortionary than wealth taxes on the living, but this benefit has not been enough to overcome administrative complexity and popular opposition to strong estate taxes.

Policy Changes and Implicit Wealth Taxes

In addition to any taxes they explicitly impose on wealth, governments also use implicit wealth taxes whenever tax policies alter the attractiveness of different assets. For example, under a switch in the individual tax base from income to consumption, investors receive a higher after-tax return on new investments but are saddled with unanticipated taxes on the decumulation of existing assets for consumption purposes (Auerbach and Kotlikoff 1987). Implicit wealth taxes of this sort are potentially much larger in magnitude than formal wealth taxes themselves.

See Also

- ▶ Estate and Inheritance Taxes
- ▶ Property Taxation
- ▶ Redistribution of Income and Wealth
- ▶ Single Tax
- ▶ Tiebout Hypothesis

Bibliography

Auerbach, A., and L. Kotlikoff. 1987. *Dynamic fiscal policy*. Cambridge: Cambridge University Press.

Carter, S., S. Gartner, M. Haines, A. Olmstead, R. Sutch, and G. Wright. 2006. *Historical statistics of the United States: Millennial edition*, vol. 5. Cambridge: Cambridge University Press.

Domar, E., and R. Musgrave. 1944. Proportional income taxation and risk-taking. *Quarterly Journal of Economics* 58: 388–422.

George, H. 1882. *Progress and poverty*. New York: Appleton.

Gordon, R. 1985. Taxation of corporate capital income: Tax revenues versus tax distortions. *Quarterly Journal of Economics* 100: 1–27.

Hicks, J., U. Hicks, and L. Rostas. 1941. *The taxation of war wealth*. Oxford: Clarendon Press.

Kaplow, L. 1994. Taxation and risk taking: A general equilibrium perspective. *National Tax Journal* 47: 789–798.

Mieszkowski, P., and G. Zodrow. 1989. Taxation and the Tiebout model: The differential effects of head taxes, taxes on land rents, and property taxes. *Journal of Economic Literature* 27: 1098–1146.

OECD (Organization for Economic Cooperation and Development). 2002. *Revenue statistics of OECD member countries 1965–2001*. Paris: OECD.

Seligman, E. 1895. *Essays in taxation*, 10th ed. New York: Macmillan, 1925.

Stiglitz, J. 1969. The effects of income, wealth and capital gains taxation on risk-taking. *Quarterly Journal of Economics* 83: 263–283.

Tiebout, C. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.

Taxes and Subsidies

J. de V. Graaff

The taxes and subsidies with which we shall be concerned are the *corrective* or *Pigovian* ones that could in theory be used to bring marginal private costs or benefits more closely into alignment with marginal social ones. The need for alignment arises when externalities (whether economies or diseconomies), operating at the margin, cause a divergence.

The name of Pigou is often associated with the idea, although his own statements are very cautious (1932, p. 381; 1947, pp. 99–100). That the matter is more complicated became clear from later work of Coase (1960), Buchanan and Stubblebine (1962) and others. A good summary is to be found in Turvey (1963).

Externalities may be between firms, where they are caused by technological interdependence between production functions, one firm's containing an input or an output proper to another's. They may be between consumers, one's utility function containing a variable proper to another consumer. Or they may, by an obvious extension, be between consumers and firms. Examples will be given later.

Optimizing behaviour by consumers and firms implies that marginal private costs and benefits will be equalized. If these coincide with social costs and benefits, they too will be equalized; if not, there will be a measure of market failure. Market failure means that all mutually beneficial bargains have not been struck and that it is possible to make one or more participants better off without making anyone worse off. Efficient markets (or 'Pareto efficiency') imply the exhaustion of all such opportunities.

The purpose of a corrective tax is to bridge the gap between private and social cost (if that is the side of the market we are looking at) created at the margin by an externality, and to vary with it. (As will shortly appear, this is much easier in theory than in practice.) The idea is to bring the externality to account, as it would be brought to account if internalized by a merger, not to eliminate it. It is to make an otherwise inefficient market simulate an efficient one, achieving by fiscal intervention what might (if transaction costs had not been too high) presumably have been achieved by direct negotiation between the parties.

A simple example (still on the cost side) will help to make these ideas more concrete. Assume that factory A discharges effluent into a river, increasing the purification costs of factory B, situated lower down, where it draws its water. Assume that B's costs of purification depend on the method chosen and vary with its product mix

and levels of output as well as with A's output and expenditure (or lack of it) on filtration. (This example is 'simple' because at least the causation runs one way only, from A to B. If both factories drew water from a lake into which both discharged effluent, reciprocal externalities could give rise to the sort of situation commonly encountered in Game Theory.)

The complexity of designing a tax on A that would correctly measure the costs imposed on B, and vary with *them* (rather then with A's activity) as A adjusted to it (by changing the quantity and quality of water discharged) and B responded, is evident. It is also evident that, to simulate the operation of a market in which there was direct negotiation between A and B, the proceeds of the tax would have to go to B, not to the fisc. This is a simple point, often overlooked, but recognized in the older literature when reference was made to systems of taxes *and subsidies*. The latter were to go not only to those creating external economies but to those suffering external costs.

From the community's point of view the most efficient (i.e. cost-effective) configuration of water treatment systems could be along the following lines. Factory A reduces its discharge of effluent (or improves its filtration) by less than it would if taxed an amount equal to the damage caused. Instead it seeks to limit the damage by sharing the costs of operating and up-grading B's purification plant. This could be *much* more efficient, offering gains to both parities. It is an outcome that could be achieved by encouraging negotiation between the two firms, or allowing them to merge. It could not be achieved by a unilateral tax.

Internalizing an externality by merger is possibly only between firms, but negotiation is possible in a wider context (between consumers, or between consumers and firms) whenever the numbers involved are not large. The that the numbers often are large. Consider motorists creating congestion on a public road. Negotiation to reduce the external costs that each imposes on the others is hardly feasible. The transaction costs would be enormous. Some sort of second best solution is to be preferred. A tax on fuel, licence fees and highway rules are among the possibilities. They would at least help to reduce congestion, or limit its consequences.

Corrective taxes are best seen in this light. They are seldom a practical way of achieving Pareto efficiency. But they could be part of a second best solution to the problem of market failure. As such they are to be weighted against other second best solutions such as licences, zoning laws and outright prohibitions. In the river pollution example, if there were too many firms to make negotiation feasible, a law prohibiting the discharge of effluent, or even one prohibiting the use of river water for industrial purposes, would clearly eliminate the externality. But it would not bring about a particularly efficient state of affairs. In such circumstances the use of admittedly rough corrective taxes might be a serious alternative.

In other circumstances a combination of measures might be appropriate. A tax on tobacco is almost always designed with an eye to revenue, the demand for the commodity being inelastic, but the acceptability of the tax is probably due to widespread recognition of the external costs imposed by smokers on non-smokers. The sheer difficulty of trying to measure these costs makes a genuinely corrective tax all but impossible. So partial prohibitions in the form of non-smoking areas are common. Their creation is a much more practical approach to a common social problem than attempting to tax smokers whenever they cause inconvenience to others – or suggesting that non-smokers should pay smokers to refrain from indulging.

This brings us to a final point. In our original example, if A and B were brought to the negotiating table, the outcome of the negotiation would be very different if the laws, customs or conventions of the society gave riparian owners (a) the right to discharge effluent or (b) the right to draw clean water. The two regimes would result in two very different distributions of the potential gains. But under either regime an efficient outcome would be possible – one in which all opportunities for mutually beneficial bargains had been exhausted. Corrective taxes and subsidies are concerned with the latter, not with the distribution of the gains from trade – that is to say: with efficiency, not equity.

See Also

► Neutral Taxation

Bibliography

Buchanan, J.S., and W.C. Stubblebine. 1962. Externality. *Economica* 26: 371–384.
Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.
Pigou, A.C. 1932. *The economics of welfare*, 4th ed. London: Macmillan.
Pigou, A.C. 1947. *A study in public finance*, 3rd ed. London: Macmillan.
Turvey, R. 1963. On divergences between social cost and private cost. *Economica* 30: 309–313.

Taylor Rules

Athanasios Orphanides

Abstract

Taylor rules are simple monetary policy rules that prescribe how a central bank should adjust its interest rate policy instrument in a systematic manner in response to developments in inflation and macroeconomic activity. This article reviews the development and characteristics of Taylor rules in relation to alternative monetary policy guides and discusses their role for positive and normative monetary policy analysis.

Keywords

Central banks; Difference rules; Econometric policy evaluation; Federal Reserve System; Forecasting; Friedman, M.; Full employment; Inflation; Inflation targeting; Interest rate rules; IS–LM models; Macroeconometric models; Monetary policy; Monetary policy rules; Monetary transmission mechanism; Money stock; Money supply; Natural growth of output; Natural rate and market rate of interest; Natural rate of employment; Nominal income; Output gap; Policy design; Price stability; Real output; Stabilization policy; Taylor rules; Time consistency; Uncertainty; Velocity of circulation; Wicksell, J. G. K

JEL Classifications

E5

Taylor rules are simple monetary policy rules that prescribe how a central bank should adjust its interest rate policy instrument in a systematic manner in response to developments in inflation and macroeconomic activity. They provide a useful framework for the analysis of historical policy and for the econometric evaluation of specific alternative strategies that a central bank can use as the basis for its interest rate decisions.

A perennial question in monetary economics has been how the monetary authority should formulate and implement its policy decisions so as to best foster ultimate policy objectives such as price stability and full employment over time. It is widely accepted that well-designed monetary policy can counteract macroeconomic disturbances and dampen cyclical fluctuations in prices and employment, thereby improving overall economic stability and welfare. In principle, when economic growth unexpectedly weakens below the economy's potential, accommodative monetary policy can stimulate aggregate demand and restore full employment. Likewise, when inflationary pressures develop, monetary restriction can restore the central bank's price stability objective. In practice, however, given the limited knowledge that economists have about the macroeconomy – for example, about macroeconomic dynamics, about the monetary transmission mechanism, and even about the measurement of fundamental concepts such as the natural rates of output, employment and interest – there is substantial disagreement about the scope of stabilization policy and about policy design.

One approach is to decide upon what seems to be the best policy on a period-by-period basis, without appeal to any specific policy guide. A seeming advantage of this approach is that it gives policymakers the discretion to use their judgement period by period. However, a basic

tenet of modern research is that systematic policy – that is, policy based on a contingency plan or policy rule – has important advantages over a purely discretionary policy approach. By committing to follow a rule, policymakers can avoid the inefficiency associated with the time-inconsistency problem that arises when policy is formulated in a discretionary manner. Following a rule allows policymakers to communicate and explain their policy actions more effectively. Policy based on a well-understood rule enhances the accountability of the central bank and improves the credibility of future policy actions. Also, by making future policy decisions more predictable, rule-based policy facilitates forecasting by financial market participants, businesses, and households, thereby reducing uncertainty.

Various proposals for monetary policy rules have been made over time, and a vast literature continues to examine the relative advantages and drawbacks of alternatives in abstract theoretical terms, in the context of empirical macroeconometric models, and in terms of the practical experience accumulated from past policy practice. To appreciate the appeal and limitations of Taylor rules, it is useful to relate their development to other proposals for systematic monetary policy.

Development of Monetary Policy Rules

Some proposals suggest postulating a rule in terms of the main objectives of monetary policy, for example 'maintain economic stability' or 'maintain a constant aggregate price level'. (See Simons, 1936, for early arguments favouring price-level targeting over discretionary policy.) One important practical difficulty with these proposals, however, is that the concepts involved are not under the control of the central bank and thus the proposals are not operational. In essence, these proposals fail to draw a clear distinction between the objectives of monetary policy and the policy instruments that are at least under the approximate control of the central bank. As a result, the suggested rules are only implicit in nature and are difficult to monitor and to distinguish from discretionary policy in a meaningful manner.

To be useful in practice, policy rules must be simple and transparent to communicate, implement and verify. This requires a clear choice of what should serve as the policy instrument – for example the money supply, m, or the short-term interest rate, i – and clear guidance as to how any other information necessary to implement the rule – for instance recent readings or forecasts of inflation and economy activity – should be used to adjust the policy instrument.

Perhaps the simplest example of a policy rule is the proposal that the central bank maintain a constant rate of growth of the money supply – Milton Friedman's k-percent rule (Friedman 1960). The rule draws on the equation of exchange expressed in growth rates:

$$\Delta m + \Delta v = \pi + \Delta q \tag{1}$$

where $\pi \equiv \Delta p$ is the rate of inflation and p, m, v, and q are (the logarithms of), respectively, the price level, money stock, money velocity, and real output. Selecting the constant growth of money, k, to correspond to the sum of a desired inflation target, π^*, and the economy's potential growth rate, Δq^*, and adjusting for any secular trend in the velocity of money, Δv^*, suggests a simple rule that can achieve, on average, the desired inflation target, π^*:

$$\Delta m = \pi^* + \Delta q^* - \Delta v^*. \tag{2}$$

Further, if the velocity of money were fairly stable this simple rule would also yield a high degree of economic stability. An early illustration of this rule appeared in 1935 in the work of Carl Snyder, a statistician at the Federal Reserve Bank of New York. After estimating that the normal rate of growth of trade in the United States was about four per cent per year at the time and observing that the velocity of money was stable, Snyder argued that 'the highest attainable degree of general industrial and economic stability will be gained by an expansion of currency and credit ... at this rate [four per cent]' (Snyder 1935, p. 198). During the 1960s and early 1970s, Milton Friedman's recommendation that the Federal Reserve control the rate of money growth to equal four per cent per year was similarly based

on the assumption that potential output growth in the Unites States roughly equalled four per cent – the prevailing estimate at that time.

Another way to interpret this policy rule is in terms of the growth of nominal income, $\Delta x = \pi + \Delta q$. With the economy's natural growth of nominal income defined as the sum of the natural growth rate of output and the central bank's inflation objective, $\Delta x^* = \pi^* + \Delta q^*$, a rule for constant money growth can be seen as targeting this natural growth rate. An advantage of a constant money growth rule is that very little information is required to implement it. If velocity does not exhibit a secular trend, the only required element for calibrating the rule is the economy's natural growth of output. In addition, while the calibration of this rule does not rest on the specification of any particular model, the rule is remarkably stable across alternative models of the economy. In this sense, the policy of maintaining a constant growth rate of money is arguably the ultimate example of a rule that is robust to possible model misspecification.

Simple modifications allowing for some automatic response of money growth to economic developments have also been proposed as simple rules that could deliver improved macroeconomic performance (see, for example, Cooper and Fischer 1972). Among the simplest such alternatives is the rule associated with Bennett McCallum (1988, 1993):

$$\Delta m = \Delta x^* - \Delta v^* - \varphi_{\Delta x}(\Delta x - \Delta x^*). \quad (3)$$

McCallum showed that, if a rule such as this (for example, with $\varphi_{\Delta x} = 0{:}5$) had been followed, the performance of the US economy likely would have been considerably better than actual performance, especially during the 1930s and 1970s – the two periods of the worst monetary policy mistakes in the history of the Federal Reserve.

A factor that complicates the use of the money stock as a policy instrument is the potential for instability in the demand for money due either to temporary disturbances or to persistent changes resulting from financial innovation. In part for this reason, central banks generally prefer to adjust monetary policy using an interest rate instrument.

A policy rule quite as simple as Friedman's *k*-percent rule cannot be formulated with an interest rate instrument. As early as Wicksell's (1898) monumental treatise on *Interest and Prices*, it was recognized that attempting to peg the short-term nominal interest rate at a fixed value does not constitute a stable policy rule. (Indeed, this was one reason why Friedman 1968, and others expressed a preference for rules with money as the policy instrument.) Wicksell argued that the central bank should aim to maintain price stability, which in theory could be achieved if the interest rate were always equal to the economy's natural rate of interest, r^*. Recognizing that the natural rate of interest is merely an abstract, unobservable concept, however, he noted: 'This does not mean that the bank ought actually to *ascertain* the natural rate before fixing their own rates of interest. That would, of course, be impracticable, and would also be quite unnecessary.' Rather, Wicksell pointed out that a simple policy rule that responded systematically to prices would be sufficient to achieve satisfactory, though imperfect, stability: 'If prices rise, the rate of interest is to be raised; and if prices fall, the rate of interest is to be lowered; and the rate of interest is henceforth to be maintained at its new level until a further movement in prices calls for a further change in one direction or the other' (Wicksell 1898, p. 189, emphasis in the original). In algebraic terms, Wicksell proposed what is arguably the simplest reactive monetary rule with an interest rate instrument:

$$\Delta i = \theta \pi. \quad (4)$$

Wicksell's simple interest rate rule did not attract much attention in policy discussions, perhaps because of its exclusive focus on price stability and lack of explicit reference to developments in real economic activity.

The Classic Taylor Rule and Its Generalizations

The policy rules that are commonly referred to as Taylor rules are simple reactive rules that adjust the interest rate policy instrument in response to

developments in both inflation and economic activity. An important advance in the development of these rules can be identified with the policy regime evaluation project reported in a volume published by the Brookings Institution (Bryant et al. 1993). The objective of the project was to identify simple reactive interest rate rules that would deliver satisfactory economic performance for price stability and economic stability across a range of competing estimated models. The Brookings project examined rules that set deviations of the short-term nominal interest rate, i, from some baseline path, i^*, in proportion to deviations of target variables z, from their targets, z^*:

$$i - i^* = \theta(z - z^*). \tag{5}$$

The collective findings pointed to two alternatives as the most promising in delivering satisfactory economic performance across models. One targeted nominal income, while the other targeted inflation and real output:

$$i - i^* = \theta_\pi(\pi - \pi^*) + \theta_q(q - q^*). \tag{6}$$

The potential usefulness of this particular rule as a benchmark for setting monetary policy was further highlighted in the celebrated contribution by John B. Taylor at the Fall 1992 Carnegie-Rochester Conference on Public Policy. Taylor developed a 'hypothetical but representative policy rule' (1993, p. 214) by using the sum of the equilibrium or natural rate of interest, r^*, and inflation, π, for i^* and setting the inflation target and equilibrium real interest equal to two and the response parameters to one half. The result was what became known as the classic Taylor rule:

$$i = 2 + \pi + \frac{1}{2}(\pi - 2) + \frac{1}{2}(q - q^*). \tag{7}$$

Taylor noted that, if one used the deviation of real quarterly output from a linear trend to measure the output gap, $(q - q^*)$, and the year-over-year rate of change of the output deflator to measure inflation, π, this parameterization appeared to describe Federal Reserve behaviour well in the late 1980s and early 1990s.

The confluence of the econometric evaluation evidence supporting the stabilization properties of this rule and its usefulness for understanding historical monetary policy in a period generally accepted as having good policy performance generated tremendous interest, and numerous central banks began to monitor this policy rule or related variants to provide guidance in policy decisions. These developments also greatly influenced monetary policy research and teaching. By linking interest rate decisions directly to inflation and economic activity, Taylor rules offered a convenient tool for studying monetary policy while abstracting from a detailed analysis of the demand for and supply of money. This allowed the development of simpler models (see the survey in Clarida et al. 1999, and papers in Taylor 1999) and the replacement of the 'LM curve' with a Taylor rule in treatments of the Hicksian IS–LM apparatus. (It should be noted, however, that this abstraction is overly simplistic when the short-term interest rate approaches zero. At the zero bound, the stance of monetary policy can no longer be measured or communicated with a short-term interest rate instrument; see, for example, Orphanides and Wieland 2000). Subsequent research (see Orphanides 2003b, for a survey) suggested that a generalized form of Taylor's classic rule could provide a useful common basis both for econometric policy evaluation across diverse families of models and for historical monetary policy analysis over a broad range of experience:

$$\begin{aligned} i = {} & (1 - \theta_i)(r^* + \pi^*) + \theta_i i_{-1} + \theta_\pi(\pi - \pi^*) \\ & + \theta_q(q - q^*) + \theta_{\Delta q}(\Delta q - \Delta q^*). \end{aligned} \tag{8}$$

The generalized Taylor rule (8) nests rule (6) as a special case but introduces two additional elements. First, it allows for inertial behaviour in setting interest rates, $\theta_i > 0$, which proves particularly important for policy analysis in models with strong expectational channels (Woodford 2003). Second, it allows the policy response to developments in economic activity to take two forms: a response to the level of the output gap, $(q - q^*)$, or its difference, which can also be restated as a response to the difference between

output growth and its potential, $(\Delta q - \Delta q^*)$. The generalized Taylor rule also nests another simplification of special interest, $\theta_i = 1$ and $\theta_q = 0$, which yields a family of difference rules similar to Wicksell's original proposal:

$$\Delta i = \theta_\pi(\pi - \pi^*) + \theta_{\Delta q}(\Delta q - \Delta q^*). \quad (9)$$

These difference rules are also of interest because, like money-growth rules, their implementation does not require estimates of the natural rate of interest or the level of potential output (and the output gap) but only of the growth rate of potential output. Indeed, these rules may be viewed as a reformulation of money-growth rules in terms of an interest rate instrument. To see the relationship of (9) to money growth targeting, note that, by substituting the money growth in rule (3) into the equation of exchange, that rule can be stated in terms of the velocity of money:

$$\Delta v - \Delta v^* = (1 + \varphi_{\Delta x})(\Delta x - \Delta x^*). \quad (10)$$

To reformulate this strategy in terms of an interest rate rule, consider the simplest formulation of money demand as a (log-) linear relationship between velocity deviations from its equilibrium and the rate of interest. In difference form this is

$$\Delta v - \Delta v^* = a\Delta i + e, \quad (11)$$

where $a > 0$ and e summarizes short-run money demand dynamics and temporary velocity disturbances. An interest-rate-based strategy that avoids the short-run velocity fluctuations, e, may be obtained by substituting the remaining part of (11) into (10). This yields

$$\Delta i = \theta((\pi - \pi^*) + (\Delta q - \Delta q^*)) \quad (12)$$

for some $\theta > 0$, which, as can be readily seen, has the same form as rule (9).

In light of this flexibility in nesting a wide range of alternative monetary policy strategies and the relative simplicity of the form (8), Taylor rules have been used to discuss a variety of policy regimes, from money growth targeting (see, for example, Clarida and Gertler 1997) to inflation targeting (see, for example, Orphanides and Williams 2007).

A crucial element for the design and operational implementation of a Taylor rule is the detailed description of its inputs. This requires specificity regarding the measures of inflation and economic activity that the policy rule should respond to, whether forecasts or recent outcomes of these variables are to be employed, and the source of these data or forecasts. In addition, the source of information and updating procedures regarding the unobservable concepts required for implementing the rule must be stipulated. Specificity in these dimensions is essential for practical analysis because there is often a multitude of competing alternatives and a lack of consensus about the appropriate concepts and sources of information that ought to be used for policy analysis. This situation is particularly vexing in regard to the treatment of unobservable concepts, such as the output gap. Unfortunately, econometric policy evaluation exercises suggest that inferences regarding the performance of a particular Taylor rule often depend sensitively on assumptions regarding the availability and reliability of these inputs. Differences in underlying assumptions complicate comparisons across studies and often explain differences in reported findings.

An illustrative example of this sensitivity relates to improper treatment of information regarding the current state of the economy. A common pitfall in theoretical policy evaluation exercises is to assume that the current state of the economy – for example, the current output gap – can be perfectly observed. Under this assumption, a Taylor rule with a vigorous response to the output gap is often recommended as 'optimal' in model-based policy evaluations. However, naive adoption of such recommendations would be counterproductive. Available real-time estimates of the output gap are imperfect, and historical experience suggests that the mismeasurement is often substantial. Under these circumstances, better stabilization outcomes would result if policy did not respond to the output gap at all or if it responded to output growth instead (Orphanides 2003a). If the natural rate of interest is also

unknown and its real-time estimates are subject to significant mismeasurement, the difference variant of the Taylor rule, (9), proves considerably more robust than the Brookings variant, (6), reversing the ranking of the two alternatives that is implied under perfect knowledge (Orphanides and Williams 2002).

Another example of such sensitivity relates to the use of forecasts in the Taylor rule. Because of lags in the monetary policy transmission mechanism, pre-emptive policy reaction is generally recommended, especially with respect to inflation. But inferences regarding the performance of forecast-based policy are sensitive to the quality of the forecasts. In some models, Taylor rules responding to several-quarters-ahead forecasts of inflation appear more promising for stabilization than rules focusing only on near-term conditions. However, this conclusion is not robust and is overturned once the potential unreliability of longer-term forecasts due to model misspecification is factored into the analysis (Levin et al. 2003).

As already noted, Taylor rules have proven valuable for historical policy analysis. Following Taylor (1993), numerous authors have examined historical monetary policy in the United States using either calibrated or estimated versions of Taylor rules (8). Studying the characteristics of policy in periods associated with good or bad economic performance helps identify aspects of policy that may be associated with such differences in performance. A complicating factor is the need for real-time data and forecasts for proper inference (Orphanides 2001). The pitfall of using *ex post* revised data and retrospective estimates of unobserved concepts in estimating Taylor rules is not uncommon. However, interpretations of historical policy based on information that was unavailable to policymakers when policy decisions were made is of questionable value. Policy prescriptions from a fixed rule are distorted as the inputs to the rule are revised from those originally available to policymakers, and therefore counterfactual comparisons of alternative policy rules can be misleading when they are based on revised data.

Despite these challenges, some useful elements of policy design emerge from historical analysis of Taylor rules, (8). First, and arguably most important, good stabilization performance is associated with a strong reaction to inflation. Second, good performance is associated with policy rules that exhibit considerable inertia. Third, a strong reaction to mismeasured output gaps has historically proven counterproductive. Fourth, successful policy could still usefully incorporate information from real economic activity by focusing on the growth rate of the economy. To be sure, such broad principles provide insufficient guidance for identifying the precise policy rule that might be ideal in a specific context. But this is not the objective of policy design with Taylor rules. Rather, the goal is the identification of simple guides that are robust to misspecification and other sources of error experienced over history.

In summary, Taylor rules offer a simple and transparent framework with which to organize the discussion of systematic monetary policy. Their adoption as a tool for policy discussions has facilitated a welcome convergence between monetary policy practice and monetary policy research and proved an important advance for both positive and normative analysis.

See Also

► Inflation Targeting

Bibliography

Bryant, R.C., P. Hooper, and C. Mann, ed. 1993. *Evaluating policy regimes: New research in empirical macroeconomics*. Washington DC: Brookings.

Clarida, R., J. Gali, and M. Gertler. 1999. The science of monetary policy. *Journal of Economic Literature* 37: 1661–1707.

Clarida, R., and M. Gertler. 1997. How the Bundesbank conducts monetary policy. In *Reducing inflation: Motivation and strategy*, ed. C. Romer and D. Romer. Chicago: University of Chicago Press.

Cooper, J.P., and S. Fischer. 1972. Simulations of monetary rules in the FRB–MIT–Penn model. *Journal of Money, Credit and Banking* 4: 384–396.

Friedman, M. 1960. *A program for monetary stability*. New York: Fordham University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.
Levin, A., V. Wieland, and J.C. Williams. 2003. The performance of forecast-based monetary policy rules under model uncertainty. *American Economic Review* 93: 622–645.
McCallum, B.T. 1988. Robustness properties of a rule for monetary policy. *Carnegie-Rochester Conference Series on Public Policy* 29: 173–203.
McCallum, B.T. 1993. Specification and analysis of a monetary policy rule for Japan. *Bank of Japan Monetary and Economic Studies* 11(2): 1–45.
Orphanides, A. 2001. Monetary policy rules based on real-time data. *American Economic Review* 91: 964–985.
Orphanides, A. 2003a. Monetary policy evaluation with noisy information. *Journal of Monetary Economics* 50: 605–631.
Orphanides, A. 2003b. Historical monetary policy analysis and the Taylor rule. *Journal of Monetary Economics* 50: 983–1022.
Orphanides, A., and V. Wieland. 2000. Efficient monetary policy design near price stability. *Journal of the Japanese and International Economies* 14: 327–365.
Orphanides, A., and J.C. Williams. 2002. Robust monetary policy rules with unknown natural rates. *Brookings Papers on Economic Activity* 2002(2): 63–145.
Orphanides, A., and J.C. Williams. 2007. Inflation targeting under imperfect knowledge. In *Monetary policy under inflation targeting*, ed. F. Mishkin and K. Schmidt-Hebbel. Santiago: Central Bank of Chile.
Simons, H.C. 1936. Rules vs authorities in monetary policy. *Journal of Political Economy* 44: 1–30.
Snyder, C. 1935. The problem of monetary and economic stability. *Quarterly Journal of Economics* 49: 173–205.
Taylor, J.B. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.
Taylor, J.B., ed. 1999. *Monetary policy rules*. Chicago: University of Chicago.
Wicksell, K. 1898. Interest and prices. Trans. R.F. Kahn. London: Macmillan, 1936.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy.* Princeton: Princeton University Press.

Taylor, Fred Manville (1855–1932)

Daniel R. Fusfeld

Keywords
Market socialism; Taylor, F. M.

JEL Classifications
B31

Taylor made his chief contribution to economic theory in his 1928 presidential address to the American Economic Association, in which he laid out the basic principles of market socialism (Taylor 1929). He argued that rational allocation of resources could be achieved in a socialist state if three conditions are met: citizens obtain income from the state in exchange for services; income is freely spent on goods offered for sale by the state at given prices; prices are set at full costs of production. The third condition can be met through a trial-and-error method in which prices of factors of production are set at levels that clear the market. Given these costs and consumer demand, markets for finished products can be cleared by adjusting levels of output and inventories. Such a system could achieve results similar to those of a competitive private enterprise economy.

Taylor's doctorate was in philosophy from the University of Michigan (1888). He taught at Albion College 1879–92, and in the Department of Economics at Michigan from 1892 to 1929. A strong advocate of laissez-faire policies and the gold standard, Taylor was a noted expositor of economic theory, with emphasis on Marshallian partial equilibrium analysis, analytic rigour and a libertarian ideology. His *Principles* textbook (Taylor 1911) went through nine editions from 1911 to 1925.

Selected Works

1911. *Principles of economics.* Ann Arbor: University of Michigan Press, 1918; New York: Ronald Press, 1921. 9th ed., 1925.
1929. The guidance of production in a socialist state. *American Economic Review* 19(1): 1–8 Reprinted in *On the economic theory of socialism,* ed. B.E. Lipincott. New York: McGraw-Hill, 1938.

Taylor, Harriet (Née Hardy, later Mrs John Stuart Mill) (1807–1858)

Josephine Kamm

Harriet Taylor was born in London on 8 October 1807 and died at Avignon on 3 November 1858. A Unitarian doctor's daughter, she was beautiful, mainly self-educated and of considerable intelligence. In 1826 she married a Unitarian wholesale druggist, John Taylor, and had two sons and one daughter, Helen, who became a champion of women's suffrage and higher education. Harriet, who had literary ambitions, contributed briefly and anonymously book reviews, verses and articles to the Unitarian *Monthly Repository.* In 1830 she met John Stuart Mill, also a contributor to the *Repository.* Friendship, based on mutual concern for the poor and the inferior status of women, developed into love. Their indiscreet public appearances and travels caused a scandal. In 1851, some two years after the death of John Taylor, who had reluctantly condoned the situation, they married. Mill praised Harriet in extravagant terms, which his friends, admirers and subsequent critics found false and ludicrous, particularly the inscription on her tombstone, ending were there but a few hearts and intellects like hers this earth would already become the hoped-for heaven.

It is only in recent years that her influence over his later work has received recognition. Its importance was apparent in *On Liberty,* published in 1859 after her death, and *The Subjection of Women* (1869) based on her only additional publication, 'The Enfranchisement of Women', which, with Mill's help and through his agency, appeared anonymously in the *Westminster Review* for July 1851. Her influence on his *Principles of Political Economy* was decisive. She persuaded him to include a chapter, 'On the Probable Futurity of the Labouring Classes', written partly in her own words. At her request he wrote, 'The poor have come out of leading-strings and cannot any longer be governed or treated like children . . .' (Mill 1848). In 1849 the chapter 'Of Property' was revised at her insistence. Initially she had approved the claim that in existing circumstances Socialism and Communism (they appear to have been considered synonymous) were neither realistic nor desirable since they overemphasized the importance of security. The 1848 French uprising which preceded the Republic had strengthened Harriet's left-wing bias. The argument, she now declared, was invalid and she took violent exception to a passage, to Mill most pertinent, that, 'The necessaries of life, when they have always been secure for the whole of life, are scarcely more a subject of consciousness or . . . happiness than the elements' (Mill 1848). He now confessed that further consideration would probably bring him to her side, as was almost invariably the case when a subject received their joint consideration. Against his former judgement he amended the passage to read, 'On the Communistic scheme, supposing it to be successful, there would be an end to all anxiety concerning the means of subsistence; and this would be much gained for human happiness . . .' (Mill 1849).

The third edition (1852) marked Harriet's triumph. If, Mill wrote, 'the choice were to be made between Communism with all its chances, and the present state of society with all its sufferings and injustices . . . all the difficulties . . . of Communism would be as dust in the balance . . . .' It was, however, too soon to judge whether, at their best, 'individual agency' or left-wing doctrines 'will be the ultimate form of human society'. Mill did not exaggerate when he wrote that, but for Harriet, the left-wing argument would 'either have been absent, or the suggestions would have been made much more timidly and in a more qualified form' (Mill, *Autobiography,* edited by Helen Taylor, 1873).

Selected Works

1851. The enfranchisement of women. *Westminster and Foreign Quarterly Review* 55.

References

Mill, J.S. 1848, 1849. Principles of political economy. In *Collected works of John Stuart Mill,* vol. III, ed. J.M. Robson. Toronto/London: University of Toronto Press/Routledge & Kegan Paul, 1965.
Mill, J.S. 1852. Principles of political economy. In *Collected works of John Stuart Mill*, vol. II, 3rd ed, ed. J.M. Robson. Toronto/London: University of Toronto Press/Routledge & Kegan Paul, 1965.
Mill, J.S. 1859. *On liberty*, 4th ed. London: Longmans & Green, 1869.
Mill, J.S. 1869. *The subjection of women*. London: Longmans & Green.
Mill, J.S. 1873. Autobiography. In *Collected works of John Stuart Mill*, vol. I, ed. J.M. Robson and J. Stillinger. Toronto/London: University of Toronto Press/Routledge & Kegan Paul, 1981.

Taylorism

A. L. Friedman

Keywords

Scientific management; Soldiering; Taylor, F. W.; Taylorism; Time study

JEL Classifications

M1

Taylorism refers to the system of management developed by Frederick Winslow Taylor. Taylor called his system 'scientific management'. Scientific management is clearly described in Taylor's two most famous works, *Shop Management* (1903) and *The Principles of Scientific Management* (1911).

Scientific management is based on the following principles:

1. Management gathers and systematizes all the workers' traditional knowledge.
2. All possible 'brainwork' is removed from the shop and centred in the planning or layout department.
3. The work should be divided into its simplest constituent elements: the tasks. Management should try to limit individual 'jobs' to a single task as far as possible.
4. Managers should specify the tasks to be done in complete detail. These tasks should be presented to the worker in written form. They should note not only what is to be done, but also how it is to be done and the exact time allowed for doing it.
5. The work should be monitored closely.

Taylor's techniques for gathering information about work was time study, the measurement of elapsed time for each component operation of a work process. Taylor also recommended that the foreman's job should be divided into more simplified task collections. Shop-floor foremen should be divided into the setting-up boss, speed boss, quantity inspector and repair boss. However, the main division was to separate work-design and manning-level decisions away from shop-floor foremen and to the planning department.

The purpose of Taylor's system was to eliminate 'soldiering', or low worker effort. This could either take the form of natural soldiering, the natural instinct and tendency for men to take it easy, or systematic soldiering, the calculated reduction of effort arising from actions and communication among groups of workers. The ultimate cause of both forms of soldiering for Taylor 'lay in the ignorance of the management as to what really constitutes a proper day's work for a workman' (1911, p. 53). Once this was determined 'scientifically', workers would be forced to comply with this standard by careful monitoring of their performance and by a differential piecework payment system. A target rate of work would be determined by work study. If workers exceeded this target, they would receive a bonus, but bonus payments would reach a ceiling between 30 per cent and 100 per cent of the standard work rate. If workers failed to meet their targets, they would lose earnings.

There is a wide range of opinions as to the importance of Taylorism. For some, Taylorism represents the dominant theory and practice of 20th-century management (Drucker 1954; Braverman 1974). For others, Taylorism is viewed as having widespread ideological impact,

but not much influence on practice because it was successfully resisted by workers and was too expensive for managers (Edwards 1979). Finally, there are those who view Taylorism as an expression of important changes in management practices, but that moves towards Taylorism have not been universal. The application of Taylorism is contingent on environmental factors (Friedman 1977; Littler 1982).

Bibliography

Braverman, H. 1974. *Labor and monopoly capital*. New York: Monthly Review Press.

Drucker, P. 1954. *The practice of management*. New York: Harper & Row.

Edwards, R. 1979. *Contested terrain*. London: Heinemann.

Friedman, A.L. 1977. *Industry and labour*. London: Macmillan.

Littler, C.R. 1982. *The development of the labour process in capitalist societies*. London: Heinemann.

Taylor, F.W. 1903. Shop management. In *Scientific management,* ed. F.W. Taylor. London: Harper & Row, 1964a.

Taylor, F.W. 1911. The principles of scientific management. In *Scientific management,* ed. F.W. Taylor. London: Harper & Row, 1964b.

Teams

Roy Radner

Abstract

A team consists of a number of decision-makers, with common interests and beliefs, but controlling different decision variables and basing their decisions on (possibly) different information. The economic theory of teams is concerned with (1) the allocation of decision variables and information among the team members, and (2) the characterization of efficient decision rules, given the allocation of tasks and information. The theory of teams thus addresses a middle ground between the theory of individual decision under uncertainty and the theory of games, and provides a natural framework for the analysis of mechanisms for decentralization, including market-like mechanisms.

Keywords

Decisions under uncertainty; Demand price; Efficient allocation; Game-playing; Games; Information costs; Informational decentralization; Optimal decision functions; Probability; Signalling; Socialism; Teams

JEL Classifications

D7

The economic theory of teams addresses a middle ground between the theory of individual decision under uncertainty and the theory of games. A *team* is made up of a number of decision-makers, with common interests and beliefs, but controlling different decision variables and basing their decisions on (possibly) different information. The theory of teams is concerned with (1) the allocation of decision variables (tasks) and information among the members of the team, and (2) the characterization of efficient decision rules, given the allocation of tasks and information.

For example, in the pre-computer age, airline companies had a number of ticket agents who were authorized to sell reservations on future flights with only partial information about what reservations had been booked by other agents. A team- theoretic issue would be the characterization of best rules for those agents to use under such circumstances, taking account of the joint probability distribution of demands for reservations at the different offices, the losses due to selling too many or too few reservations in total, and so forth. A second issue would be the calculation of the increase in expected profit that would be obtained by providing each agent with better information about the status of reservations at all offices, or by increased centralization of the reservation process. To calculate this increase in expected profit – the *value* of the additional information – one of course needs to know something about the best decision rules with and without the additional information. However, providing this additional information would require

additional communication, transmission, processing, and storage, all of which would be costly. The value of the information puts an upper bound on the additional cost that should be incurred. The value – and cost – of the information will depend on its structure and on the structure of the team's decision problem, and not just on some simple measure of the 'quantity' of information. (For a study of the airline reservation problem, and other models of sales organization, from a team-theoretic point of view, see Beckmann 1958 and McGuire 1961, respectively.)

In this entry we shall sketch a formal model of team theory, the characterization of optimal team decision functions, and the evaluation of information in a team. The theory will be illustrated with a discussion of decentralized resource allocation, followed by concluding remarks on the incentive problem.

A Formal Model

We consider a *team* and *M members.* Each member m controls an *action,* say a_m. The resulting utility to the team depends on the *team action*,

$$a = (a_1, \ldots, a_M),$$

and on the state of the environment. (Since the team members have common interests, there is a single utility for the whole team.) The state of the environment comprises all the variables about which team members may be uncertain before choosing their actions. It is determined exogenously, i.e. is not subject to the control or influence of the team members. If we denote the state of the environment by s, then we can denote the utility to the team by u(a, s); the function u will be called the payoff function for the team.

Before choosing an action, each team member m receives an information signal, y_m. This information signal is determined by the state of the environment, say $y_m = \eta_m(s)$. (This includes the case of 'noisy' information, if the description of the state of the environment includes a description of the noise.) We shall call η_m the *information* function for member m, and the M-tuple $\eta = (\eta_1, \ldots, \eta_M)$ will be called the *information structure* of the team.

Each team member m will choose his action on the basis of the information signal he receives, according to a decision function, say α_m. Thus

$$a_m = \alpha_m(y_m) = \alpha_m(\eta_m[s]). \tag{1}$$

If we use the symbol α to denote the *team decision* function, i.e., the M-tuple of individual decision functions, then the utility to the team, in *state s*, of using the information structure η and decision function α can be expressed as

$$U(s) = u[\alpha(\eta[s]), s], \tag{2}$$

To express the team's uncertainty about the state of the environment, we suppose that s is determined according to some probability distribution, φ, on the set S of possible states. This probability distribution may be interpreted as 'objective' or 'personal'; in the latter case it represents the beliefs of the team members (Savage 1954). It is part of the definition of a team that its members have common beliefs, as well as common utility functions.

With the state s distributed according to the probability distribution φ, the utility $U(s)$ in (2) is a random variable. We shall assume that the team chooses its decision function so as to maximize the (mathematical) expectation of this utility,

$$E[U(s)] = \sum_s \varphi(s)U(s) \equiv \omega(\alpha, \eta, \varphi). \tag{3}$$

As a special case, suppose that the information functions of the team members are identical. In this case, the team decision problem is formally identical to a one- person decision problem in which the same person controls all of the actions. An alternative interpretation of this case is that the information is *centralized.* By contrast, if at least two team members have essentially different information functions, then we may say that the information is *decentralized*. With this definition of (informational) decentralization, we see that all organizations but the very smallest are likely to be decentralized to some extent.

The expected utility for the team depends on the team members' decision function, the team information structure, and the probability distribution of states of the environment, as well as on the 'structure' of the decision problem, i.e. the way in which the utility (2) depends on the members' action and the state of the environment. This is brought out by the notation in (3). If we want to compare the usefulness of two different information structures, we have to associate with them some corresponding decision function α and probability distribution φ. Since the probability distribution (sometimes called the 'prior distribution') represents either objective probabilities or the team members' common beliefs before they receive further information, it is natural to take it as a datum of the problem. On the other hand, since the decision functions can be chosen by the team, it is natural to associate with each information structure the corresponding team decision function that maximizes its expected utility (given the information structure). Thus the optimization problem for the team may be posed in two stages: (1) for a given information structure, characterize the optimal team decision function(s); (2) optimize the information structure, taking account of the costs of – or constraints on – making the information available, and with the proviso that for each information structure the team uses an optimal decision function.

More will be said below about each of these stages of the problem. However, it should be emphasized here that the choice of information structure comprises most of the organizational design choices that are not concerned with conflicts of interests or beliefs among the organization's members. The information structure is, of course, affected by the pattern of observation and communication in the team. In addition, the allocation of tasks within the team is expressed by the information structure. To see this, suppose that each member of the team were assigned an information structure; then a reassignment of decision variables to team members would be formally equivalent to reassigning information functions to decision variables.

Optimal Decision Functions

We shall now consider the characterization of team decision functions that are optimal for a given structure of information. It will be useful to recall here the corresponding problem for a single-person decision problem (see, e.g., Marschak and Radner 1972, ch. 2). We may use the same model and notation as in the previous section, but remembering that there is only one member of the team. The following statement provides a general characterization of the optimal decision function: *For each information signal, choose an action that maximizes the conditional expected utility given the particular signal.*

This characterization is easily derived from Eqs. 2 and 3. In Eq. 3, group the terms in the sum according to the information signal associated with each state; this gives us

$$E[U(s)] = \sum_{y} \sum_{\eta(s)=y} \varphi(s)U(s). \tag{4}$$

From (2), if $\eta(s) = y$, then the resulting utility in that state is

$$U(s) = u[\alpha(y), s]. \tag{5}$$

For each signal y, the decision-maker can choose an action $a = \alpha(y)$. Hence, combining (4) and (5) we see that, for each signal y, the decision-maker should choose $a = \alpha(y)$ to maximize

$$\sum_{\eta(s) \ =y} \varphi(s)u(a, s). \tag{6}$$

Let $\psi(y)$ denote the probability of y, and let $\varphi(s|y)$ denote the conditional probability of s given y. By definition,

$$\psi(y) = \sum_{y=\eta(s)} \varphi(s),$$

and if $y = \eta(s)$,

$$\varphi(s|\, y) = \frac{\varphi(s)}{\psi(y)},$$

or

$$\varphi(s) = \psi(y)\varphi(s|\, y).$$

Hence (6) can be written as

$$\psi(y) \sum_{\eta(s)=y} \varphi(s|\, y)u(a,s), \tag{7}$$

so that maximizing (6) is equivalent to maximizing

$$\sum_{\eta(s)=y} \varphi(s|\, y)u(a,s), \tag{8}$$

which we recognize as the conditional expected utility using the action a, given the signal y. This proves the above characterization of the best decision function. (Notice that we have implicitly assumed that the signal has positive probability. There is no loss of generality in doing so; we can simply exclude from consideration all signals that have zero probability, since they do not affect the expected utility.)

The characterization of optimal single-person decision functions can be extended to the case of a team, but in a restricted way. Consider a particular team member i. If a team decision function, say $\hat{\alpha}$, is optimal, then surely i's decision function α_i is optimal given that each other member j uses $\hat{\alpha}_j$. Hence i is faced, so-to-speak, with a one person decision problem in which the other members' decision functions form part of i's 'environment'. The following is therefore a *necessary* condition for a team decision function to be optimal:

Person-by-Person-Optimality Condition

For each member i, and for each signal y_i with positive probability, the corresponding action $a_i = a_i(y_i)$ maximizes the team's conditional expected utility given the signal, y_t and the decision functions of the other members.

Although person-by-person-optimality is necessary for optimality, it need not be sufficient. However, one can prove the following:

Theorem 1 If each member's action is a real finite-dimensional vector chosen from some open rectangle, and if for each state s the team's utility is a concave and differentiable function of the team action, then any team decision function that is person-by-person-optimal is also optimal. (For a proof of this theorem, and an example in which a person-by-person-optimal decision function fails to be optimal, see Marschak and Radner 1972, ch. 10, s. 3.; for a more complete treatment, see Radner 1962.)

The person-by-person-optimality condition can be applied to yield more detailed characterizations of optimal team decision functions for special cases, e.g., in which the utility function is quadratic or piecewise-linear (see Marschak and Radner 1972, ch. 10). A few such applications are illustrated below.

The Evaluation of Information

As noted at the beginning of this discussion, many of the most interesting qsts in organizational design concern the comparison of alternative information structures. One information structure is better than another to the extent that it permits better decisions; on the other hand, this improvement can be obtained only at some additional cost.

We first consider the case in which the utility from the decisions is additively separable from the cost of the information structure; we shall call this the *separable case.* In this case, one is justified in defining the *gross value* of an information structure as the difference between (1) the expected utility derived from its best use and (2) the maximum utility obtainable using no information (beyond that contained in the prior probability distribution of states).

If the team has no information (the null information structure), then its decision function reduces to a single team action. The maximum expected utility that the team can obtain with the null information structure is

$$V_0(\varphi) = \max_a \sum_s u(a,s). \tag{9}$$

Hence in the separable case, the gross value of an information structure η is defined as

$$V(\eta, \varphi) \equiv \max_{\alpha} \omega(\alpha, \eta, \varphi) - V_0(\varphi). \qquad (10)$$

(Cf. Eq. 3.)

Note that the value of an information structure depends on the prior distribution, as well as on the entire structure of the decision problem (available actions, utility function, etc.). This should make one suspect that there is no way to tell whether one information structure is more valuable than another just by examining the two information structures alone.

To examine this question more carefully, it is useful to introduce another representation of information. Consider again for the moment the single-person case; an information structure then consists of a single function from states to information signals. For any given signal, there is a set of states that give rise to that signal. This correspondence between signals and sets of states determines a partition of the set of states; each element of the partition is a set of all states that lead to a particular signal; denote this partition by (S_y). It is obvious that any two information structures that give rise to the same partition are equivalent from the point of view of the decision-maker, and in particular must have the same value. In other words, the names or labels of the signals are unimportant.

Suppose now that the set S of states of the environment, the number M of team members, and the set A of team actions are fixed. Consider the family of all team decision problems that can be formulated with the given triple (S, M, A). In other words, consider the set of all pairs (u, φ), where u is a utility function for the team, and φ is a prior distribution, compatible with (S, M, A). We shall say that one information structure *is as valuable* as another information structure if the value of the first is greater than or equal to the value of the second for all team decision problems compatible with (S, M, A). (Value is defined by Eq. 10.)

The following criterion provides a simple test for the relation 'as valuable as'. Of two partitions of the set S, we shall say the first is as *fine* as the second if every element of the first partition is a subset of some element of the second (the first can be obtained by 'refining' the second). Let $\eta = (\eta_1, \ldots \eta_M)$ and $\chi = (\chi_1, \ldots, \chi_M)$ be two team information structures. We shall say that η is as fine as χ if for every team member m, η_m is as fine as χ_m. One can prove (see Marschak and Radner 1972, ch. 2, Sec. 6):

Theorem 2 Assume that every team member has at least three alternative actions; then the information structure η is at least as valuable as the information structure χ if and only if, for each member m, η_m is as fine as χ_m.

Theorem 2 can be extended to deal with 'noisy' information (see, e.g., McGuire 1972).

Since two partitions of a set need not be ranked by the relation 'as fine as', it is clear from Theorem 2 that the relation 'as valuable as' is only a partial ordering of information structures. This implies, in particular, that there is no numerical measure of 'quantity of information' that can rank all information structures in order of value, independent of the decision problem in which the information is used.

If the utility of the team decision and the cost of information are not additively separable, then an alternative definition of value of information must be used. For example, suppose that the outcome of the team decision and the cost of the information structure are both measured in dollars, and the team is not riskneutral, so that the team utility is some (nonlinear) function of the outcome and the cost. Then we can define the value of the information structure as the 'demand price', i.e., the smallest cost that would make the team indifferent between using the information structure and having no information beyond the prior distribution. (For further discussion of the comparison of information structures, see McGuire 1972. For more on the value of and demand for information, see Arrow 1972.)

Decentralization

We have used the term informational decentralization to refer to a structure of information in which not all members have the same information

function. In an economic organization the information structure is generated by processes of observation, communication, storage, and computation. For example, suppose that each team member m starts by observing a different random variable, say $\zeta_m(s)$. If there were no communication among the members before actions were taken, then each member's information would be the same as his observation – an extreme form of decentralization. On the other hand, if there were complete communication of their observations among the members, then their information functions would be identical, namely $\zeta = (\zeta_1,\ldots, \zeta_M)$. Alternatively, the latter information structure could be generated by having all members communicate their observations to a central agency, which would then compute the team action and communicate the corresponding individual action to each member. In the last two cases, we would say that the information structure is completely centralized, because all of the members' actions were based on the same information.

Rarely does one encounter in a real organization the extremes of no communication or complete communication just described. Rather, one finds that numerous devices are used to bring about a partial exchange of information. The usefulness of such devices is measured by the excess of additional value (expected utility) they contribute over the costs of installing and operating them. Examples of such devices are the dissemination of reports and instructions, the formation of committees and task forces, and 'management by exception'. Formal models and a comparative analysis of some of these devices are given in (Marschak and Radner 1972, ch. 6). In particular, this methodology is used to elucidate the value of two different forms of management by exception.

Allocation of Resources in a Team

For many economists, the purely competitive market represents the ideal model of economic decentralization. Indeed, in some economic literature, 'decentralization' and 'pure competition' are synonymous. The potential usefulness of market-like mechanisms to decentralize economic decision-making in a socialist economy has also been discussed by students of socialism (Lange and Taylor 1938; Lerner 1944; Ward 1967).

The theory of teams provides a natural framework for the analysis of market mechanisms as a device for decentralization. For example, consider the problem of allocating resources to productive enterprises. Suppose that some resources are initially held centrally by a 'resource manager'. Before any exchange of information, the resource manager observes the supplies of centrally available resources, and each enterprise manager observes his respective local conditions of production: technology, supplies of local resources, etc. The action of the resource manager is to allocate the central resources among the enterprises. The action of an enterprise manager includes (say) the choice of techniques and the levels of inputs of local resources. The state of the environment comprises the total supplies of central resources and the local conditions.

At one extreme, the team action could be taken without any communication. In particular, the central resources would be allocated based only on the prior probability distribution of local conditions. Regarding the supplies of central resources, each enterprise manager would know only the prior probability distribution of such supplies, and the allocation rule to be used by the resource manager. We might call the resulting information structure 'routine'.

At the other extreme (complete centralization), each enterprise manager might be required to report to the resource manager all of his information about local conditions. The resource manager would then compute both the optimal decisions of the enterprises and the optimal allocation of resources. Accordingly, the resources would be allocated and the enterprise managers would then be 'instructed' by the resource manager as to what actions they should take.

In a market mechanism, the resource manager would announce prices (of central resources), and the enterprise managers would respond with demands. In the literature on allocation and price-adjustment mechanisms it is usually assumed that this exchange of messages is iterated until an equilibrium of supply and demand is

reached (this may require infinitely many iterations!). In a real application of such a mechanism, only a few iterations would typically be feasible, and equilibrium would not be reached. Thus one could not appeal to the theory of optimality of the equilibria of such processes. Nevertheless, the exchange of information produced by even a few iterations might be quite valuable, i.e., the information structure might be much more valuable than the 'routine' structure, and possibly close in value to that of complete centralization.

Indeed, research done to-date on models of such processes suggests that price and demand signals are strikingly efficient in conveying the information needed for good allocation decisions, even out of equilibrium (Radner 1972; Groves and Radner 1972; Marschak 1959, 1972; Arrow and Radner 1979; Groves and Hart 1982; Groves 1983; Hogan 1971).

Incentives in Teams

The model of a team assumes that the team members have identical interests and beliefs. Thus no special incentives are required to persuade the individual members to honestly implement the given information structure or to take the decisions prescribed by the optimal team decision function. A full-fledged theory of economic organization should, of course, take account of conflicting interests and beliefs, and the resulting problems of incentives.

This article is not the place to review the growing literature on this subject, but a few comments may be useful here. In general, it is not possible to solve the 'incentive problem' costlessly. (For exceptions to this generalization, see Groves 1973; Green and Laffont 1979.) Thus, in an economic organization, there will be two sources of efficiency loss: (1) decentralization of information, having the effect that individual actions will be based on information that is less complete than the information jointly available to the organization as a whole; (2) conflicts of interests and beliefs among the decision-makers, leading to distortions of information and action ('gameplaying'). In fact these two sources are not so easily disentangled. For example, under conditions of uncertainty and limited information, it will typically be difficult for a supervisor (or organizer) to determine whether a particular decision-maker is providing correct information or following a prescribed decision rule, since to achieve this would require the supervisor to have all of the information that is available to the subordinate. In other words, informational decentralization leads to de facto decentralization of authority. (For references to the literature on incentives and decentralization in economic organizations see Arrow 1974; Hurwicz 1979; Radner 1975, 1986; and Stiglitz 1983.)

See Also

- ▶ Efficient Allocation
- ▶ Signalling and Screening

Bibliography

Arrow, K.J. 1972a. The value of and demand for information. Ch. 6 of McGuire and Radner (1986).

Arrow, T.A. 1972b. Computation in organizations: The comparison of price mechanisms and other adjustment processes. Ch. 12 of McGuire and Radner (1986), 237–282.

Arrow, K.J. 1974. *The limits of organization*. New York: Norton.

Arrow, K.J., and R. Radner. 1979. Allocation of resources in large teams. *Econometrica* 47: 361–385.

Beckmann, M.J. 1958. Decision and team problems in airline reservations. *Econometrica* 26: 134–145.

Green, J., and J.-J. Laffont. 1979. *Incentives in public decision-making*. Amsterdam: North-Holland.

Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.

Groves, T. 1983. The usefulness of demand forecasts for team resource allocation in a stochastic environment. *Review of Economic Studies* 50: 555–571.

Groves, T., and S. Hart. 1982. Efficiency of resource allocation by uninformed demand. *Econometrica* 50: 1453–1482.

Groves, T., and R. Radner. 1972. Allocation of resources in a team. *Journal of Economic Theory* 3: 415–444.

Hogan, T.M. 1971. A comparison of information structures and convergence properties of several multisector economic planning procedures. Technical Report No. 10. Center for Research in Management Science, University of California, Berkeley.

Hurwicz, L. 1979. On the interaction between information and incentives in organizations. In *Communication and*

control in society, ed. K. Krittendorf, 123–147. New York: Gordon and Breach.
Lange, O., and F.M. Taylor. 1938. *On the economic theory of socialism*. Minneapolis: University of Minnesota Press.
Lerner, A. 1944. *The economics of control*. New York: Macmillan.
Marschak, T.A. 1959. Centralization and decentralization in economic organizations. *Econometrica* 27: 399–430.
Marschak, J., and R. Radner. 1972. *Economic theory of teams*. New Haven: Yale University Press.
McGuire, C.B. 1961. Some team models of a sales organization. *Management Science* 7: 101–130.
McGuire, C.B. 1972. Comparisons of information structures. Ch. 5 of McGuire and Radner (1986), 101–130.
McGuire, C.B., and R. Radner. 1986. *Decision and organization*. 2nd ed. Minneapolis: University of Minnesota Press. originally published Amsterdam: North-Holland, 1972.
Radner, R. 1962. Team decision problems. *Annals of Mathematical Statistics* 33: 857–881.
Radner, R. 1972. Allocation of a scarce resource under uncertainty: An example of a team. Ch. 11 of McGuire and Radner (1986), 217–236.
Radner, R. 1975. Economic planning under uncertainty. In Ch. 4 of *Economic planning, east and west,* ed. M. Bornstein, 93–118. Cambridge, MA: Ballinger.
Radner, R. 1987. Decentralization and incentives. In *Information, incentives, and economic mechanisms: Essays in hornor of Leonid Hurwicz*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
Stiglitz, J.E. 1983. Risk, incentives, and the pure theory of moral hazard. *The Geneva Papers on Risk and Insurance* 8: 4–33.
Ward, B. 1967. *The socialist economy*. New York: Random House.

Technical Change

S. Metcalfe

Abstract

Successive transformations of economic society from agricultural to industrial form and beyond to the service economy have consolidated a process of economic change with its own inner logic of tremendous power, a logic which harnessed continual technical and organizational innovation to the pursuit of profit. The intertwining of emergent knowledge and economic adaptation within the instituted frame of modern capitalism is at the core of economic self-transformation. This article treats this topic in three parts: the relation between new knowledge and economic transformation; some consequences of technical change; and the residual productivity debate.

Keywords

Additivity; Adjustment costs; Capital intensity; Capitalism; Capital–labour substitution; Consumption growth frontier; Diffusion of technology; Division of labour; Economic growth; Efficient allocation; Entrepreneurship; Equilibrium; Industrial Revolution; Innovation; Input–Output analysis; Internet; Invention; Knowledge; Labour productivity; Learning; Leisure; Market institutions; Maturity; Population growth; Production functions; Productivity growth; Research and development; Residual productivity; Roundabout methods of production; Specialization; Standards of living; Structural change; Technical change; Thrift; Total factor productivity; Trigger effects; Wage–Profit frontier

JEL Classifications

O3

The successive transformations of economic society since the 14th century from agricultural to industrial form and beyond to the service economy have consolidated a process of economic change with its own inner logic of tremendous power, a logic which harnessed continual technical and organizational innovation to the pursuit of profit. At the core of the new logic is the intertwining of emergent knowledge and economic adaptation to its hidden possibilities that has been the basis for the sustained increase of aggregate output per person employed – the chief proximate source of increased standards of living in the Western world – the progressive mechanization and automation of production methods, and

the continuous development of the economic structure (Kuznets 1977; Mokyr 1990, 2002). The gains in material welfare, in length of human life, in life experience and functioning have been beyond anything achieved before the 18th century, yet knowledge-driven progress comes at a price. It is necessarily uneven in its incidence across space and time, and the ensuing disparities of performance can and do impose heavy human adjustment costs as old ways give ground to the new. Skills are devalued, capital assets lose the capacity to generate income, while there is little prospect of the losers receiving compensation. If the balance sheet speaks to progress, it does so in a tangled way. This is the ethic of competition, and nowhere is this unevenness more apparent than in the seemingly unavoidable differences in economic performance between advanced and developing countries. Knowledge-driven economic growth is never a smooth, balanced affair of proportional expansion with each activity advancing in step. Rather, as Schumpeter insisted, it involves disharmony and fierce competition between new and old activities and places, a diversity of growth rates and profit rates and continual reallocation of labour and capital between and within activities. It is occasionally useful to study such processes as if structural change were absent, but to do so courts the danger of missing the substance and therefore the process and significance of technical and economic change. For the very process of uneven adjustment is a powerful stimulus to the development of further knowledge. This is the Faustian bargain accepted by the Western world with its origins in Reformation and Renaissance and its consolidation in the 19th and 20th centuries.

What was the nature of this emergent combination of knowledge-generating system and market system of economic adaptation? All productive activity involves the transformation of materials and energy, in a purposeful, intelligent and information-dependent fashion to add value to the materials and energy. These transformations are of physical form, or of spatial location, or of availability in time; indeed, the history of technical change is a history of invention and innovation directed to providing new inanimate energy sources, to providing means to control the application of energy, and to providing new, synthetic materials on which to work. What economists see from one perspective as the substitution of capital for labour in more roundabout methods of production is from this perspective the substitution of non-human for human energy, a process that was well established through the use of water power in the early middle ages and went on to be revolutionized by innovations in steam power, internal combustion and ultimately electricity. All existing industries were affected by the new power sources, (textiles, iron and steel, mining in particular) in the first Industrial Revolution, but new industries emerged, too, to produce the machines of increasing sophistication and specialization to harness the new power sources and to extend their application to transport and communication. That much product innovation was required to exploit the new possibilities warns us that the relation between product and process technology is usually close. That new forms of business organization were needed to deal with increased capital intensity and new risks tells us that technical and organizational change were seldom far removed from each other. The increasing incentives to exploit the material base of the planet and the ability to synthesise materials not found as natural compounds gave further licence to the innovation process with oil, aluminum, plastics and pharmaceuticals, each becoming commonplace in the 20th century. But it would be a mistake to focus attention on the great, traditional industries alone, whether producing capital goods, consumer goods or intermediate products; trends of a different nature reflected the deeper ways in which new knowledge was working its transformative effects. Superficially this is seen in terms of the displacement of manual labour as an energy supply, but this misses the point that the human role was moving increasingly into one of the management and coordination of information that is essential for any rational activity. Within business enterprises, within public bureaucracies and within markets, greater human effort was required to manage and coordinate the flows of information demanded by economic growth and its handmaiden – a richer and deeper division of

labour. That this could be possible only in the presence of a productivity-enhancing revolution in information generating, storing and communicating activities should be obvious at least from our position in the early 21st century. In this regard, the innovation sequence and economic adaptations associated with the printing press proved to be of profound importance: for it eliminated a long-standing constraint on the exact reproduction of information, and made possible its transmission over generations (storage), and its more ready transport. Written communication increased in relative importance compared with face-to-face conversation, the costs associated with codification declined dramatically and, consequentially, the organization of a spatially distributed but deeply connected process of the growth of science and technology became possible (Eisenstein 1979). The subsequent developments of telegraph and telephone, of information-processing machines, of television and radio, and now of the Internet are amplifications of the revolution begun by printing. Since all societies are knowledge based, that epithet counts for little: what matters is the form of information society that prevails, and the way in which inanimate energy and its harnessing machinery has been applied progressively to further transform the production, transport and storage of information is a development at least as significant as the discovery of economical steam power in Watt's day. The growth in the productivity of information activities made possible in manufacturing, transport and communication, and now service activities, is accepted by all commentators as immense, for it has economized on scarce mental capacity, not just on limited human strength. Quite remarkably, total employment has continued to rise even though its occupational composition, like the composition of economic output, is continually shifting in response to new technological possibilities, despite the prognostications of less confident observers of modern capitalism. Modern capitalism appears, by accident no doubt, to have evolved a set of institutional rules which not only promotes the efficient use of what is known to further specific human ends but serves also to greatly stimulate the production of further new knowledge. The solution of one problem through the speculative deployment of imagination serves only to open up further problems, while the future direction and outcomes of this process remain necessarily hidden from view. Knowledge and economy are deeply intertwined, and the restless nature of the one reinforces the restless nature of the other. This is the nature of that Faustian bargain between a knowledge system and the market process.

As the above sketch might warn the reader, the economic analysis of technical progress is not a straightforward matter. The familiar tools of equilibrium economics are best suited to discussing the long-run effects of new products and methods of production; they are not well suited to analysis of the disequilibrium processes by which new technologies are generated, improved and absorbed into the economic structure. All these processes take time, operate with different velocities and are subject to complicated interactions and feedback effects. It has been traditional to divide the analysis of technical change into three branches: *invention*, the creation of new products and processes; *innovation*, the transfer of invention to commercial application; and *diffusion*, the spread of innovation into the economic environment. Unfortunately, this has provided a somewhat fragmented approach to the study of technical change in which an understanding of interdependence and feedback between the stages has been hidden, together with important elements which emphasize the continuity of advance.

A more unified approach is possible if we place the growth of knowledge and its instantiation in novel innovations in the context of a competitive market process, in which firms seek to differentiate themselves and gain commercial advantages by introducing new products/services and processes. That the market system is a defining instituted feature of modern capitalism goes a long way to explaining its experimental proclivities and thus the importance to it of enterprise. It is an open system in which, in principle, all existing constellations of resources are open to challenge by rival, entrepreneurial conjectures according to their anticipated profitability. Innovations (like breakthroughs in science) are always statements

of disagreement, in this case about the efficacy of the prevailing allocation of resources. Thus there is an important boundary in play, to work the system needs order and agreement, to progress it needs disorder and disagreement, it is the institutions of capitalism that contribute to keeping the two in balance. Of course, firms are not the only important source of new knowledge; they are only one element in a more comprehensive system of public and private research institutes of many kinds; but with respect to innovation (as distinct from invention) they play a virtually dominant role as the one combinatorial agency bringing together the different kinds of problem solving and exploiting resources needed to create wealth from knowledge. The way in which they combine internal with external innovation resources is an important aspect of their innovative ability, as Marshall knew well; external organization, connecting to customers, suppliers and other sources of knowledge, is as much a part of a firm's productive capital as is its internal organization. The three questions which arise from this viewpoint are: (*a*) by what processes is technological variety generated?; (*b*) by what processes do different varieties acquire economic weight?; and, (*c*) by what mechanisms does the process of acquiring economic weight shape the development of technological variety?

To the first question belongs the study of a firm's strategy in changing its knowledge base and articulating new and improved products and processes. The role of science in modern invention, the organization of R&D activity, and the links between a firm and other knowledge-generating institutions each play a role with respect to variety and its generation. But no individual variety of product or process is significant until it acquires economic weight, and the greater the weight the greater the impacts of the new technology upon its environment. In regard to the second question, innovations acquire economic significance because they are superior either from the point of view of users or from the point of view of their producers or both. Clearly, however, the more profitable it is to use new products and processes and the more profitable it is to supply them, the more quickly will they acquire economic weight and displace existing products and processes. The dynamics of adjustment to new opportunities depend on how different the new technologies are from established forms, and how the economic environment evaluates those differences relative to the standards of value and cost. The third question contains some of the most complex questions of all, relating to the inducement mechanisms which generate and shape technological variety. Clearly there are important non-economic factors at work. However, different environments for market exploitation do make it profitable to develop a technology in different directions, and the experience of exploiting a technology in a given environment, more often than not, gives rise to important learning effects which indicate an agenda for subsequent development and applications in other areas of activity. Technologies do not emerge into the economic sphere fully fledged but typically in immature form, and evolve very much according to the bottlenecks and incentives to development which arise in their application. Progress thus tends to be localized around a canalized path of advance and to be contingent on factor prices and the values consumers place on different combinations of functional features. The same technological opportunity exploited in different environments would in all probability develop in different directions. By a similar token, a technology which is mature in one environment may be developing rapidly in another. Maturity is at root an economic concept applicable to situations where the expected benefits fall short of the expected costs of further advancing technology.

Within the development of economic thought, the study of technical change has never played a major role. Indeed, from Adam Smith onwards, and with the exceptions of Marx and Marshall, it was progressively written out of classical economic analysis. Thus, despite Smith's emphasis on the division of labour as a form of induced technological and organizational change, little of his remarkably productive insight survived in subsequent writings, apart from the maintained separation of the agricultural and manufacturing sectors as different loci of progress and increasing returns. Not surprisingly, no classical writer

foresaw that technical progress in agricultural methods would dispel the niggardliness of nature and banish the spectre of the stationary state. By the time Robbins came to write his methodological characterization of the neoclassical scheme in 1932, not only had technical progress been handed over to the psychologists and engineers, but the very nature of the questions posed by economists had changed fundamentally. Gone was the emphasis on accumulation and progress and in its place stood the analysis of the allocation of given resources under given technical conditions and, moreover, subject to a definition of competition as a state of equilibrium quite incompatible with the increasing-returns implications of the division of labour. The analysis of an organic process became instead the search for the solution of a given jigsaw puzzle. Only Schumpeter (1911) provided a clear way forward. He insisted that technical progress be viewed as a transformation arising from *within* the capitalist system, that it was an integral part of the competitive process and that a key role was played by the entrepreneur and entrepreneurial profits in the process by which technologies acquire economic weight. Orthodox equilibrium theory, it will be noted, had found no room for the entrepreneur. It has been left to the post-1945 generation of economists to reassert the importance of technological change. So far they have done so in a piecemeal, empirical fashion with little attempt to reintegrate the phenomena back into a formal framework of accumulation and structural change. The writings of Pasinetti (1981) and of Nelson and Winter (for example, 1982) can be said, from quite different perspectives, to make this step and have stimulated many others to follow their lead and develop an evolutionary approach to technical change (Dosi 2000; Nelson and Winter 2002; Witt 2003).

Some Consequences of Technical Change

If the process of technical change remains difficult to handle, we can still make limited progress with an analysis of its consequences using long-period methods of analysis. Here, one of the most compelling features of production in modern industrial societies is its roundabout nature. The Industrial Revolution placed modern economies on a path of increasingly roundabout production arrangements in which resources are devoted to elaborate chains of production where raw materials (mineral or agricultural) are worked into intermediate commodities for further processing into final commodities and services with the aid of complex tools and machinery. Specialization and the division of labour are the natural features of such roundabout, mechanized methods, as Adam Smith made clear. When discussing technical progress it is particularly important to recognize that the majority of changes occur within the structure of input–output relations and not only in the activities producing final commodities for consumption (Pasinetti 1981). A framework which treats this structure as a black box into which primary inputs flow and final outputs emerge will not be a useful foundation for the study of technical progress and its effects.

To illustrate some possibilities we employ the following analytic device. Consider a self-contained component of an economic system; we call it a subsystem, which produces a single consumption good, cloth, via three separate, constant returns to scale activities. A lathe is produced with inputs of labour and itself, a loom is produced with inputs of labour and the lathe, and final output, cloth, is produced with labour and the loom. The lathe and the loom are produced means of production; they are outputs of one activity and inputs into another productive activity (Kurz and Salvadori 1995). Imagine this subsystem to be embedded in a competitive capitalist economy, and that it is analysed in long-period equilibrium conditions in which capital invested in each activity in the subsystem supports a common rate of profits, r, and grows at the common rate, g. There is no structural change taking place in the relative importance of the three activities. Given the profits rate, there will be a unique pattern of relative production prices of the three commodities and a unique level of the real wage, w (ratio of money wage to the price of cloth). Similarly, given the growth rate there is a unique pattern of employment within the subsystem and a

unique level of consumption per worker, c (ratio of cloth output to total employment in the subsystem). Now it is well known that higher values of r are related to lower values of w, while higher values of g are related to lower values of c. The corresponding so-called wage–profit and consumption growth frontiers are downward sloping, satisfy the dual property that $r = g$ when $w = c$, and have a common, finite maximum value for r and g, corresponding to zero w and zero c, respectively. These frontiers are a convenient vehicle with which to explore the effects of technical change.

Starting from a position in which only one production process is available in each activity, consider the long-run equilibrium effects of technical change. Two basic categories of change may be considered, in each case involving changes in one or more input–output coefficients in the subsystem: first, improvements, which imply no qualitative change to any output or input and require only that less of at least one existing input is used within at least one of the processes; and second, inventions, which do imply qualitative change, a physically different output (for example, a new lathe or loom) is produced by an entirely new process.

Whatever the precise changes in input–output coefficients, inventions and improvements can always be classified into three groups by comparing the long-period properties of the new bundle of processes with those of the existing bundle. Dominant technical changes are those which are economically superior over the entire range of profit rates consistent with the existing technology. At the ruling real wage and relative price structure associated with the 'old' method, the new process supports a higher rate of profit, and this is the basis for its superiority and – we must here conjecture – its adoption in the subsystem. By similar reasoning, redundant technical changes are those which are economically inferior for all possible wage and price constellations; they constitute failed inventions. Finally, conditional changes are those whose superiority or otherwise do depend on the prevailing relative price structure. For the invention or improvement to become an innovation, it must be economically superior when evaluated at the prevailing price structure. Only dominant changes and the superior set of conditional changes satisfy this condition and can have an economic effect – that is, become innovations.

The long-period effects of innovations depend on the nature of the change in technology and the position of the corresponding process in the input–output structure. In particular, changes in the machine producing processes have quite different consequences from changes within the cloth activity. The more important consequences may be summarized as follows. An improvement or invention in a machine process will alter the entire relative price structure of the subsystem. At the ruling rate of profits, the price of the commodity whose method is improved is reduced relative to the price of all other produced commodities, while the price of all commodities which use the output of the improved process are reduced relative to the money wage. The further down the chain of input–output relations lay the improvements, the greater is the breadth of the consequences of the technical change. Consequently, the simplest case involves an improvement to the cloth activity: cloth falls in price but the relative prices of lathes and looms are unaffected. The corollary of these effects is that any technical change increases the real wage consistent with the ruling rate of profits. Corresponding to the changes in price relations are changes in the structure of employment within the subsystem. A labour-saving technical improvement in a machine activity reduces the proportion of total subsystem employment absorbed by that activity, but how it redistributes employment among the other activities depends on the particular nature and location of the change in question. Any improvement in cloth making, by contrast, has no effects on the equilibrium employment structure. All technical changes will increase the level of consumption per head consistent with the ruling growth rate. Naturally, the magnitude of these effects depends on the ruling values of the growth rate and rate of profits. It is often convenient to summarize the effects of technical change in terms of the associated differences in w–r and c–g frontiers before and after the technical

change. In brief, all dominant changes give rise to new frontiers which lie above the ones associated with the old methods. In the case of conditional changes, the old and new frontiers intersect at least once. Nothing clear-cut can be established about the effects of technical change on the aggregate degree of mechanization, whether measured by the value capital–labour or the value capital–output ratio. Depending on the basis of valuing the capital stock, capital intensity may increase or decrease and the different measures may even move in opposite directions. The concept of neutral technical progress has traditionally been a focus of attention in relation to the effects of progress upon the distribution of income. As an example, the traditional case of Harrod-neutral technical progress (no effect on the value capital–output ratio at the ruling r and g) is achieved, trivially, with an improvement in labour productivity confined to the final consumption activity but, more generally, requires that labour productivity increase in equal proportionate amounts in each and every process. Such Harrod-neutral changes leave the structure of employment and relative commodity prices unchanged. There is little doubt that neutral progress of any kind is not to be expected in practice, nor is it a particularly interesting analytic category. Indeed, at given r and g values, Hicks's neutral technical progress (no effect on the capital–labour ratio) is logically impossible in an input–output subsystem of the kind discussed here (Steedman 1985). More interesting in terms of technological interdependence are the induced technical changes known as trigger effects (Simon 1951; Fujimoto 1983). Where technical progress occurs in a machine activity, it may so alter the relative profitability of other activities in which that good is an input that it becomes profitable to adopt different processes within those other activities. In this way the effects of technical change in any machine activity may trigger changes in production methods far beyond the activity in question.

In summary, even under the hypotheses of long-run equilibrium conditions the consequences of technical progress are complex, and are associated with changes in relative prices, real incomes and physical patterns of employment of all inputs. Unless attention is confined to progress in consumption goods, the full ramifications of technical progress can be understood only within an input–output framework. A fortiori one can only understand the inducements to change technology within such a framework of technological interdependence.

The Residual Debate

A central focus for the literature on technical progress has been provided since the early 1950s by a debate on the measurement of total factor productivity and the implications which follow for our understanding of the growth process. Within the neoclassical tradition, the sources of economic expansion were considered to be population growth and thrift, with growth in labour productivity dependent upon the substitution of capital for labour. Despite the early protests of Schumpeter (1911) that these mechanisms were of negligible significance in explaining long-term growth of capitalist economies, it was not until a series of studies demonstrated the apparent independence of output growth from accumulation that debate could be engaged. The ingenious methods of Abramovitz (1956), Solow (1957) and Kendrick (1973) showed beyond reasonable doubt that the modern growth of the US economy was in proportionate terms at least three-quarters due to increased efficiency in the use of productive inputs and not to the growth in the quantity of resource inputs per se. The implication was quite devastating: an adequate explanation of economic growth appeared to lie outside the traditional concerns of economists, to constitute a residual hypothesis.

From these early studies followed a lengthy sequence of extensions and amendments (that continues unabated today) creating a rich tapestry of data on the growth of the major industrialized and developing nations, and their constituent activities. For our purposes it is the framework employed to identify the contributing sources of economic growth which is of primary interest. For the measured quantities of inputs and outputs are

brittle constructs easily swayed by errors of measurement or aggregation, and particularly marked by a failure to allow for quality change in consumption and capital goods, the disamenities of modern growth and the valuation to be placed on enhanced leisure time. Despite the sophisticated efforts to refine measures of the productive input, taking detailed account of the effects of education on labour quality (Denison 1962) and on the measurement of capital goods and their services (Jorgenson and Griliches 1967), agreement on the size of the so-called residual element in growth remains as elusive as ever. Here lies a paradox: the importance of new skills and of inventions and quality improvements in capital and intermediate goods increases with the rate of technical progress, as the effects of the information revolution confirm; thus the faster the rate of progress the more the difficulty in measuring its contribution to economic growth. The rise of the service economy and its intangible outputs and inputs certainly adds to these difficulties, and it has become commonplace to suggest that the increased importance of services will reduce the possibilities for further growth in total factor productivity, let alone its accurate measurement (Griliches 1992). Certainly these considerations increase the merits of attempts to measure productivity growth at the level of more finely defined activities and take advantage of new micro datasets (Bartlesmann and Doms 2000). We cannot explore this further here, other than to point to the fact that aggregate productivity growth is now to be treated as a combination of improvements within activities and the structural changes that reallocate output and resource inputs between activities. Productivity change in this frame takes on a more evolutionary hue, as sketched above and premised on the unevenness of progress and adaptation to it (Nelson 1989).

The central organizing concept behind the early studies was the aggregate production function and the separation of observed growth in output per worker into two independent and additive elements: capital–labour substitution, reflected in movements around a given production function; and increased efficiency in resource use, as reflected by shifts in this function. To maintain additivity, the analysis had to be confined to marginal changes in output and input, and could not be applied cumulatively to longer periods without introducing an interaction term between capital substitution and increased efficiency. Within this framework all inputs, the factor services, stand on an equal footing, and constant returns alloyed with universal perfect competition allow marginal productivity pricing to identify the contribution which the growth or relative decline of each input makes to the growth of output per worker. To identify the growth of total factor productivity in a short time interval one need only subtract from the growth of output the growth in total factor input, itself a factor-price-weighted sum of the growth rates of the individual inputs. The sensitivity of such a procedure to errors of measurement in inputs, outputs and relative prices will be obvious.

Some difficulties, immediately apparent from the controversy over capital and distribution, now enter the picture. From the point of view of the long-run supply of productive services, all inputs do not stand on an equal footing. In particular, the flow of capital services depends on the stocks of usable capital instruments and thus on the ability of the economic system to maintain and augment such stocks in quality and quantity. But these capital instruments are produced and reproduced by productive activities which themselves are subject to technical progress over time. Thus, to treat independently increases in efficiency and increases in the stock of capital goods is at least misleading, unless one maintains that technical progress occurs only in consumption activities. The consequences of this for the measurement of total factor productivity are severe (Rymes 1971). To illustrate, consider an economy growing at a constant rate over time with a constant saving ratio, with the rate of increase in labour productivity the same in all activities, and the capital–output ratio constant. In such an economy the rate of increase in efficiency due to technical progress is exactly measured by the rate of increase in productivity per worker, and not by the measured increase of total factor productivity; which is of a smaller magnitude since it wrongly deducts the effects of induced capital–labour substitution. Increased efficiency makes it easier to reproduce

capital goods, such that all the observed rate of increase in the capital–labour ratio (equal to the growth of output per worker) is induced by the enhanced efficiency in the processes producing capital goods. There is no independent capital deepening to contribute to the growth of labour productivity. It is not surprising that when we identify labour as the only primary input then the natural measure of increased efficiency is the rate of increase of labour productivity. Capital goods are after all instruments made by labour too, indeed in some traditions of thought they are described and analysed as so much 'stored-up labour'. All this, of course, leaves untouched a second aspect of the capital controversy, namely, the severe conditions which have to be imposed to generate an aggregate production function along which output per worker is positively associated with the quantity of capital per worker, and for which input prices may be claimed to measure the corresponding marginal products of factor services (Bliss 1975; Harcourt 1972). It is perhaps for this reason that studies of residual productivity have become more prominent at the industry level with as detailed a specification as possible of the relevant physical flows of factor services. But disaggregation does not avoid the problem and the fact that the capital inputs of one activity are derived from the outputs of other activities. The growth of labour productivity in any one activity depends not only upon its own increase in efficiency but upon increased efficiency in the activities supplying it with capital goods, materials and energy. Thus we are back in the world of input–output interdependence in which the results of enhanced efficiency are imported and exported between activities in the way outlined in the previous section.

There can be no doubt as to the value of the residual productivity debate; it awakened interest in the origins and effects of technical progress and stimulated several new lines of research. However, it never did attempt to answer the questions about the constitution and generation of residual productivity growth. These remain the dominant questions as we seek to further understand the complex mechanisms that link technical change to the growth of wealth in modern capitalism.

See Also

- ▶ Creative Destruction
- ▶ Neo-Ricardian Economics
- ▶ Sraffian Economics
- ▶ Structural Change

Bibliography

Abramovitz, M. 1956. Resource and output trends in the United States since 1870. *American Economic Review, Papers and Proceedings* 46: 5–23.

Bartlesmann, E., and M. Doms. 2000. Understanding productivity: Lessons from longitudinal data. *Journal of Economic Literature* 38: 569–594.

Bliss, C.J. 1975. *Capital theory and the distribution of income*. Amsterdam: North- Holland.

Denison, E.F. 1962. *The sources of economic growth in the United States*. New York: Committee for Economic Development.

Dosi, G. 2000. *Innovation, organisation and economic dynamics*. Cheltenham: Edward Elgar.

Eisenstein, E. 1979. *The printing press as an agent of change*. Cambridge: Cambridge University Press.

Fujimoto, T. 1983. Inventions and technical change: A curiosum. *Manchester School of Economics and Social Studies* 51: 16–20.

Griliches, Z., ed. 1992. *Output measurement in the service sectors*. Chicago: Chicago University Press.

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Harrod, R. 1948. *Towards a dynamic economics*. London: Macmillan.

Jorgenson, D., and Z. Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34: 249–283.

Kendrick, J. 1973. *Postwar productivity trends in the United States, 1948–1969*. New York: NBER.

Kurz, H., and N. Salvadori. 1995. *Theory of production: A long-period analysis*. Cambridge: Cambridge University Press.

Kuznets, S. 1977. Two centuries of American economic growth: Reflections on US experience. *American Economic Review, Papers and Proceedings* 67: 1–14.

Mokyr, J. 1990. *The lever of riches*. Oxford: Oxford University Press.

Mokyr, J. 2002. *The gifts of Athena*. Oxford: Oxford University Press.

Nelson, R.R. 1989. Industry growth accounts and production functions when techniques are idiosyncratic. *Journal of Economic Behavior & Organization* 11: 323–341.

Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA/London: Belknap Press.

Nelson, R.R., and S. Winter. 2002. Evolutionary theorizing in economics. *Journal of Economic Perspectives* 16 (2): 23–46.
Pasinetti, L.L. 1981. *Structural change and economic growth*. Cambridge: Cambridge University Press.
Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.
Rymes, T. 1971. *On concepts of capital and technical change*. Cambridge: Cambridge University Press.
Schumpeter, J. 1911. *The theory of economic development*. Oxford: Oxford University Press. 1934.
Simon, H. 1951. Effects of technical change in a linear model. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York/London: Wiley/Chapman & Hall.
Solow, R. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39: 312–320.
Steedman, I. 1985. On the 'impossibility' of Hicks' neutral technical progress. *Economic Journal* 95: 746–758.
Witt, U. 2003. *The evolving economy: Essays on the evolutionary approach to economics*. Cheltenham: Edward Elgar.

Technology

Joel Mokyr

Abstract

Technology is the utilization of natural phenomena and regularities for human purposes. Propositional knowledge – sets of statements about natural regularities and phenomena – provides the epistemic base of technology, whose width largely determines society's ability to generate technology. The high social rate of return to technology leads to underinvestment in new technology, justifying some government subsidization. Only since the eighteenth century have producers and scientists been systematically linked, allowing technology to flourish. Technology is the main factor driving economic growth; the scope for technology transfer ensures it will continue to be so, as long as the appropriate institutions are in place.

Keywords

Ancient Rome; China; Competences; Economic growth; Endogenous technology; Enlightenment; Industrial revolution; Innovation and invention; Institutions; Intellectual property rights; Islam; Isoquant; Japan; Judaism; Patents; Production function; Propositional knowledge; Rate of return; Research; Science; Techniques; Technology

JEL Classification

N1; O3

Technology may be defined as the utilization of natural phenomena and regularities for human purposes. It is a matter of debate whether the manipulation of regularities in human behaviour by firm management should be properly classified as technology, or whether technology should be defined more narrowly as the harnessing of physics, chemistry and biology. Either way, it has become a distinctive feature of *homo sapiens*, and the success of the species in expanding technology has had momentous consequences for our planet, both physically and biologically.

In standard neoclassical economics, technology is regarded as a mapping from inputs to outputs. However, such definitions inevitably assign technology to a 'black box' category (Rosenberg 1994). Yet the economic approach illustrates the many aspects of technology relevant to social science. Much of this is captured in the concept of the isoquant, which is a summary description of the mapping. The isoquant displays the three important economic aspects of technology. One is the basic constraint that human knowledge imposes on what people can do. The lowest isoquant – that is, the fewest inputs that combine to produce a unit of output – denotes that limitation to human knowledge at any given moment. There is no implication that a more efficient production is not *possible* in some metaphysical sense, but rather that conditional on the state of knowledge in time t this is the best that can be done. Second, the isoquant map indicates that not all producers are necessarily

producing at best-practice technology. The entire set above the lowest isoquant is feasible, and while these techniques are by definition less efficient than best practice, there may be many good reasons why average practice is often considerably below best practice. Finally, the fact that the isoquant is a curve and not a point indicates one of the fundamental features of technology, namely, that there are many ways to skin a cat. One of the deepest issues in economics is the choice of technique from the available menu and how that choice is affected by economic parameters. The shape of the isoquant, moreover, tells us a great deal about the nature of the technology available, the degree to which factors are substitutes for one another, the rate at which the marginal products of factors are declining, and so on.

The production function approach only implicitly allows recognition of technology's fundamental nature, namely, that it is, above all, knowledge. By writing the function to include a shift factor, we allow for the growth of knowledge to enable an economy to do things it could not do previously. That technology is first and foremost human knowledge is not always fully recognized. Needless to say, in order to result in production, this knowledge in the vast majority of cases requires some strongly complementary inputs (tools, materials and energy), which economists define as capital and intermediate inputs. Yet in the deepest sense technology exists in the knowledge defining how certain actions plus certain inputs lead to outcomes we deem desirable.

For that reason, the fundamental unit of technology can be regarded as the *technique*, a concept close to Nelson and Winter's 'routine' (Nelson and Winter 1982). A technique is basically a set of instructions on how to produce, much like a simple recipe. Some techniques may, of course, be hugely complicated, with many conditional and nested statements, but their syntax remains prescriptive. Hence the term 'prescriptive' knowledge, which contains anything from baking a cake to driving instructions to engineering handbooks. The master-set of all techniques available in society is what Joan Robinson (1956) once called 'the book of blueprints' and it constitutes a monstrous menu from which firms and economists make selections. Two types of questions about this menu suggest themselves: how do agents really learn the contents of the menu and make selections, and how did the menu get written in the first place?

The full meaning of technology can be realized by adding the concept of *competence*. Competence concerns the execution of the instructions in the technique by agents. Instructions can be codified (in writing or orally), but no set of instructions is ever complete: they need to be read, interpreted, and carried out. If it were possible to write a *complete* set of instructions that would be wholly self-contained and self-explanatory, competence would be irrelevant and production could be entirely carried out by automatons. It is clear, however that all techniques contain implicit or 'tacit' components that require the agency of a human to interpret and carry out the instructions. The size of the 'tacit' component varies over time and from field to field, but can never be reduced to zero (Cowan and Foray 1997). It consists of a certain savoir-faire that comes with experience or imitation, but is hard to learn from codified information.

Technology and Knowledge

To understand why and how technology contains what it does at any given time, we need to consider more carefully where it comes from. Many standard definitions of technology refer to 'science' as an essential ingredient. Thus the Oxford English Dictionary defines technology as '(1). the application of scientific knowledge for practical purposes. (2). the branch of knowledge concerned with applied sciences'. Such a definition is patently ahistorical: the close association between 'science' (in the modern sense of a consensual, formal, and analytical understanding of natural phenomena) and technology is a product of the past two centuries. For many centuries, people had been employing technology in a variety of fields, yet it is hard to think of a medieval blacksmith, a peasant in biblical Palestine, or a miner in

ancient Roman Spain as relying on 'science'. Technology is therefore part of production in whatever form we observe; science came into the picture only very recently.

As an alternative to the somewhat anachronistic emphasis on science, I have proposed for historical purposes the concept of *propositional knowledge* (Mokyr 2002). Propositional knowledge is a set of statements about natural regularities and phenomena. These may be expressed in terms of firm regularities such as the laws of thermodynamics or purely in terms of description, measurement, and cataloguing. The distinction between propositional and prescriptive knowledge seems obvious: the planet Neptune and the structure of DNA were not 'invented'; they were there prior to discovery, whether we knew it or not. The same cannot be said about diesel engines or aspartame. Polanyi (1962, p. 175) notes that the distinction is recognized by patent law, which permits the patenting of inventions (additions to prescriptive knowledge) but not of discoveries (additions to propositional knowledge). He points out that the difference boils down to the observation that prescriptive knowledge can be 'right or wrong' whereas 'action can only be successful or unsuccessful'.

The main point is that this knowledge supports the prescriptive knowledge that is the essence of technology. This support is the *epistemic base* of technology. This base can be narrow or wide, depending on how much of the natural regularities of the technique is known. But it was common for things to be invented despite a narrow or negligible epistemic base. Through luck and serendipity, through dogged trial-and-error, or through an intuitive sense that defies precise analysis, inventors stumbled upon things that worked and worked well, without actually understanding *why and how* they worked. Such concepts are of course relative. It may seem to us, for example, that Alessandro Volta, who built the first working electrical battery in 1800, did not know quite how and why his 'pile' worked, but our own understanding of this, while broader than his, may still be quite limited compared with what may be known about the subatomic nature of electricity in the future.

The width of the epistemic base determines to a great extent the effectiveness of the process whereby society creates new technology. Hit-and-miss experimentation or a 'try-every-bottle-on-the-shelf' method may well yield new techniques that work, but they will tend to be one-off advances, which soon enough reach the upper bound of their capacity. Further adaptation and tweaking following an invention is far more effective if the basic modus operandi is understood. Moreover, if one does not understand why something works, it will be hard to know what does not work. Enormous amounts of human energy were misallocated, largely by highly talented individuals, in research on alchemy, astrology, attempts to build *perpetuum mobile* machines, and similar impossibilities. Advances in propositional knowledge eventually terminated these programmes. In other words, the lack of propositional knowledge greatly increased the costs of research and development, and until about 1750 most new technical advances soon ran into diminishing returns in terms of their further improvement and development. Lack of an adequate epistemic base also often curtailed the effectiveness of existing technology. For instance, in agriculture, the knowledge that fertilizer increased yields had existed for 1000 of years but, until nineteenth-century organic chemistry widened the epistemic base, basic distinctions between nitrates, phosphorus and potassium were not made, and thus often enough the quantities and kinds of fertilizers used were poor. Better understanding allowed these to be calibrated exactly, which brought about huge improvements in yields.

Technology as an Economic Good

Much of modern growth theory relies on 'endogenous technology' in which technology is being produced by inputs within the economy and responds to prices and costs (Aghion and Howitt 1997). This literature has successfully dealt with many of the uncomfortable characteristics of technology in models of a growing economy. Of those, a number stand out. One is that like all forms of knowledge it is a purely non-rivalrous

good. By giving it away for free, the original owner loses no knowledge of his own, but his capability to exploit this knowledge for commercial ends is reduced. Moreover, by giving it away he loses control over its diffusion since it is normally hard to prevent the new owner from giving it to a third. Second, new knowledge is the ultimate example of increasing returns, a good that is fixed and of no marginal cost to produce. Third, much technology is inappropriable in the sense that, once one person has it, others can often easily imitate or reverse-engineer the technique. Fourth, techniques are hard to exactly quantify, since they often have complex relations with other techniques, ranging from the purely complementary to the pure substitute. Hence, attempts to somehow 'count' the number of techniques in a society and to relate them to inputs seems ill-fated and to violate Einstein's dictum that some things that count cannot be counted and a more axiomatic way of measuring technology is needed (Olsson 2000). Fifth, the process of technology generation is subject to far more uncertainty than any other economic activity. Moreover, this is uninsurable risk. Each invention is made only once, so that there is only a limited amount one can learn from the experience of other inventions. The risk is not only that the technique an inventor is trying to write may not be feasible (or at least not feasible for her), it is also that even if the search is successful someone else may have got there first or the technique may not be commercially exploitable (Rosenberg 1996). Finally, much of the underlying propositional knowledge, the foundation and essential input of inventive activity, is available at no charge from scientific literature. Yet accessing it may be rather costly all the same.

The production of new technology has changed dramatically over past centuries. Until late in the nineteenth century, the lone inventor slaving away in a small workshop or lab was the paradigmatic creator of new technology. Some of the more successful ones may have worked on commission (such as the great Richard Roberts, the foremost mechanical engineer in Britain in the first half of the nineteenth century), but basically they were individuals working on their own account. Some of them were professional inventors who hoped – often in vain – to find the 'killer ap' that would make them rich. Others did not bother about the money, and made their inventions for their own satisfaction or for the benefit of mankind, and demonstratively refused to take out patents. Since the late nineteenth century an increasing share of inventive activity has become part of 'corporate R&D', an organized and often bureaucratized form of activity, often systematic and always driven by a corporate bottom line. The corporate research lab first emerged in the big German chemical concerns in the late nineteenth century, but the system was soon adopted in other industrialized nations. The individual inventor working in the proverbial garage has not been quite eliminated, and even today it happens that some lone wolf will come up with an idea that the huge research labs of large corporations did not think of. The agility and creativity of the single human mind may not altogether perish in the dark-suited world of corporate profits, but it is also clear that, when such an invention is successful, the road to riches usually leads to control by or a merger with a larger company with access to credit, marketing networks, and development facilities.

Despite the many market failures in the knowledge industry, however, there is a market for technology since it is clearly valuable (Arora et al. 2004). One form this market takes is technical consulting, through which firms purchase expertise not otherwise available to them. This practice can be easily traced back to the early eighteenth century, and by the time the Industrial Revolution came along consulting engineers of a variety of kinds was common among technologically advanced firms. Another way the market for technology operated was through licensing. Licences were bought and sold commonly in countries in which the patent system worked effectively to protect intellectual property rights, and they are as close as we can get to seeing how technological knowledge was valued (Khan and Sokoloff 2001). Patent protection of one form or another was quite common in eighteenth-and nineteenth-century America and Europe, and firms could buy technology owned by another firm, a practice that has continued into our own

time. Once a patent expires, however, the technique becomes common access. For that and other reasons, some industries elected to protect their techniques by secrecy, of which the best-known example is the still secret recipe for Coca Cola, code-named 'Merchandise 7X', kept under lock and key in a vault in the Sun Trust Bank Building in Atlanta, Georgia.

From a social point of view, it still seems to be the consensus that most societies seriously underinvest in the creation of new technology. This largely reflects the high social rate of return, which is widely regarded to be higher than the private rate of return resulting from the difficulty of appropriating and exploiting all the benefits of new technology and spillovers from one industry to another as well as across different countries (Mansfield et al. 1977). Such rates of return are notoriously hard to compute, and differ substantially among industries, to say nothing of the difficulty in distinguishing between average and marginal rates. But overall these rates are significantly higher in innovation than in other investment projects. The production of 'knowledge' is thus widely regarded as a market requiring some form of government intervention through the subsidization of pure scientific research and the support of some technologies with high rates of spillover.

Technology as a Historical Force

Simple models that relate complex social systems to a single technological advance or to a number of them have been proposed by some historians, but have not found a large following. These models include Lynn White's suggestion that the feudal system followed from the adoption of the horse stirrup, and Karl Wittfogel's notion that oriental despotisms had their origins in hydraulic technology and the need to coordinate water control (Smith and Marx 1994). Economists, too, have felt that at times technological developments did affect economic performance and that certain key inventions such as printing with moveable fonts or navigational techniques developed in the fifteenth century did have major effects on history, although in most cases they are non-committal about what the real exogenous variable is. What is missing from the historical record before 1750 is much evidence that technology was ever powerful enough to bring about sustained economic growth such as to make a significant difference in living standards within a reasonable time. This is not to say that the episodes of high technological creativity in Song China or Renaissance Europe did not lead to major qualitative changes in the control that people had over their resources or their daily lives. But, in so far as they led to economic growth, the effects were limited in time and space, and were often offset by population responses.

Beyond that it should be noted that most societies that ever existed were not technologically creative (Mokyr 1990). Even societies that can take credit for substantial cultural or artistic achievements, such as Greek, Hellenistic, or Jewish societies, were often not terribly inventive. Indeed, to take a global look at human history, the miracle is that technology actually changed as much as it did. Until quite recently, inventors were widely regarded as dangerous. This was in part because every act of invention is in some sense an act of rebellion and disrespect towards earlier generations and their know-how. For societies that held the wisdom of their forefathers in deep respect, such as Judaism and late Islam, inventors were little different from heretics. Moreover, inventions often threatened to reduce the value of existing human or physical capital by making it obsolete and in some cases redundant. Many governments saw disequilibrium caused by technological shocks as a threat to the status quo, and took a 'make-no-waves' approach toward new technology. Entrenched interests often took a Luddite attitude towards innovation, blocking it where they could. While Enlightenment Europe started to challenge every conventional wisdom and embarked on a new path, three literate and sophisticated empires – the Ottomans, Ch'ing China, and Tokugawa Japan – each in its own way closed off most innovation and chose stasis over progress. Second, in most societies, there was a deep social gap between, on the one hand, educated and informed people who studied nature

and mathematics and, on the other, those who did the grunt work in the fields, mines, or workshops. Many improvements that were seemingly within reach of Roman society, such as casting iron and eyeglasses, were not achieved. The conventions and social class structure that prevented this kind of communication were slowly bridged in medieval Europe by monks, who were simultaneously the educated class and deeply interested in applying new technology such as windmills and mechanical clocks. But not until eighteenth-century Europe was there an organized and concerted effort to bring those who knew things and those who made things in direct contact with one another. Only after that happened could producers access and use the propositional knowledge of the natural philosophers (as they were called then), while at the same time the needs of manufacturers and farmers began to affect the research agendas of those in charge of expanding knowledge.

Modern Science and Technology

After 1820 or so, the connections between science and technology become slowly tighter, but it is unclear which of the two was the dog and which the tail. The relation between the two varied substantially from industry to industry and from technique to technique. In some areas the science came first and then informed the technology (which in turn may then have led to a further sophistication of the science). This was surely true in a field like telegraphy, in which Oersted's famous discovery of electromagnetism in 1819 led scientists to speculate that an electromagnetic telegraph was possible. It was equally true in medicine after 1870 when scientists demonstrated that bacteria were the cause of many infectious diseases, leading to a series of techniques in preventive hygiene. But in other areas the technology came first and science followed, often at a distance. Consider materials technology: steel had been made for centuries before some of Lavoisier's students finally realized in 1786 what gave it its special properties. Even after that, the great breakthroughs in steel-making of the 1850s and 1860s were only marginally informed by the metallurgy of the time. In energy technology the gap was even larger: it took the world almost a century and a half after Newcomen's first successful engine in 1712 to finally nail down the principles that made it work.

Any simple statement about the sequencing of science or technology is in any case likely to be false. The two complemented and reinforced one another, theory and practice working cheek by jowl. Technology often operated as a 'focusing device' for scientists, showing them a well-defined problem they could then try to solve. In many cases, technology confirmed or inspired theoretical work. Heinrich Hertz's work on oscillating sparks in the 1880s and the subsequent development of wireless communications by Oliver Lodge confirmed Maxwell's purely theoretical work on electromagnetic fields. The success of the Wright brothers at Kitty Hawk in 1903 resolved the dispute among physicists on whether heavier-than-air machines were feasible at all. Following their successful flight, Ludwig Prandtl published his magisterial work on how to compute airplane lift and drag using rigorous methods.

The simple 'linear model' in which pure science leads to applied science and from there to technology is further undermined by the important feedback from technology to science that has been called 'artificial revelation'. In many fields science has been constrained by technology: astronomy depended on telescopes, microbiology on microscopes, chemistry on electrical batteries. In our own time, fast computers have become an indispensable tool for virtually every field of research. In many cases, significant scientific progress occurs when the tools to measure, to observe, or to analyse were significantly improved (Price 1984). In that way, it could be argued, technology has become self-reinforcing, and the historical models in which technology shocks have no persistence and eventually asymptote off to a new equilibrium have become irrelevant. This is particularly true because modern technology increasingly has the capability to combine and hybridize techniques with other, seemingly unrelated, techniques. Some techniques, indeed, have had such strong and so many complementarities with others that they have been dubbed 'general purpose

technologies' – steam, electricity, steel, and lasers all come to mind (Bresnahan and Trajtenberg 1995). Such hybridizing technology can grow at a dazzling pace, even if no pure new knowledge is added, simply through a growing number of combinations and recombinations (Weitzman 1996). Moreover, modern communications and access technology make it much easier for inventors to scan what is available and find the 'right' match to create ever more sophisticated hybrids.

Technology and Growth

The exponential growth in technological capabilities is responsible for the emergence of modern growth. However, not all growth derives from improved technology: improved allocations and scale economies account for some proportion of it. But growth based on knowledge is different in some important respects from other forms of growth. One is that it seems much harder to reverse. Whereas wealth based on the gains from trade can be quickly lost due to political turmoil or war, knowledge is much like the proverbial genie that cannot be placed back in the bottle. Although there are historical cases of knowledge actually being 'lost', they tend to be rare and in a modern economy quite hard to imagine. More controversial is the question of whether the accumulation of knowledge will ever run into diminishing returns through the exhaustion of technological opportunities or the proliferation of knowledge beyond our capability to contain and control it even with the very best access technology. Concerns that 'everything that can be invented has been invented' have been made repeatedly in the past and been held up to ridicule as often by historians of technology. Indeed, technological progress often requires more and better new technology simply because many techniques have unforeseen consequences that require modification or replacement. Internal combustion engines were one of the defining inventions of the twentieth century, but their impact on the environment has increasingly emphasized the need for an alternative approach. Similar instances of technological 'bite-back' can be observed in a host of other modern techniques and, while not *all* of them necessarily have a technological 'fix', better knowledge surely is part of the solution to any problems caused by new techniques. Similarly, technological successes create new needs, which themselves create an endogenous demand for new techniques. Thus the unprecedented increase in life expectancy requires an entire new set of techniques catering to the needs and wishes of people in advanced age brackets, who were a negligible proportion of the population only a century ago.

What is striking about the history of technology is that change has as often as not been competence-reducing rather than competence-increasing. Much effort has gone into making modern technology easy to operate and maintain, with the ingenuity being frontloaded in the design. Once the design is perfected, it can be mass-produced and operated by workers of relatively low skill, and increasingly by automatons, a process Marx termed 'deskilling'. While this is surely not true across the board, it is increasingly the case not just in manufacturing but in services as well. Such routinization of technology suggests a possible bifurcation in the demands for competence. On the one hand, new techniques will be devised by highly skilled scientists and engineers, whose access to the appropriate propositional knowledge is almost immediate, and whose ingenuity will drive continuous progress. However, their numbers are sufficiently small that their supply is not really a serious constraint. They can be drawn from the elite applicants to the top technological universities in the world, picked by their mathematical skills and creativity. It is a small elite of original, skilled, and driven minds that drives technological progress, as it always has. The fact that their designs can be implemented and maintained by workers whose skills may be quite limited means that the progress of technology is not really constrained much by the supply of human capital. Technology transfer is a concept that captures the very real possibility that knowledge invented in one society can in the end be utilized effectively in another, if the institutional parameters are properly lined up. There seems, hence, little reason to believe that technology-driven growth is a temporary phenomenon.

Technology and Institutions

Technology is the key to worldwide economic growth, though it is clear that in many Third World nations the lack of good institutions stands in the way of the adoption of more productive new technology. The payoff structure, both for the generation of new and for the adoption of existing technology, is determined by institutions, and this sets the stage for technological outcomes. The analysis of economists has tried to segment growth between technological achievements that push the product possibility frontier out and institutional achievements that move the economy closer to that frontier. Yet the interactions between the two make such decompositions hazardous. Economists have long realized that inventors, just like everybody else, respond to incentives. So, for that matter, do the natural philosophers and mathematicians who provide them with the epistemic base for their new techniques. Yet the exact nature of the proper payoff for those who add to the stock of useful knowledge is the subject of some debate.

Whatever the rewards for successful invention, society needs above all to ensure that those who experiment and research are not penalized, even if their research appears absurd or offensive to most others. Penalizing people because their ideas are eccentric or 'heretical' has become rather rare in our age but, with the intensification in the resistance of such ideological organizations as animal-rights or anti-nuclear groups, and the rise of public concern about sensitive areas such as human cloning and stem cells, certain fields of research may be in jeopardy. In the absence of sticks, what the optimal carrots are is far from agreed upon. The most widely used reward is patents, but historically patents have been quite ambiguous as a tool to encourage technological progress (Jaffe and Lerner 2004). The alternatives to patents all have advantages and drawbacks. Secrecy, of course, is the most costly since the social marginal costs of sharing information are zero. Moreover, any system of intellectual property rights based on secrecy discriminates against those techniques that lend themselves to reverse-engineering or obvious imitation. Awarding the three p's (prizes, pensions, patronage) to successful inventors has coexisted in many places with the fourth (patents). In *ancien régime* France, the Royal Academy was authorized to award such distinctions to inventions that benefited the realm. That such decisions were highly subjective at best and open to nepotism and corruption at worst seems obvious, but no creative society has ever been able to avoid them: Samuel Crompton and Edmund Cartwright, two of the most successful inventors of the British Industrial Revolution, were voted substantial awards by the British Parliament because they had failed to secure patent protection. The South Carolina legislature awarded Eli Whitney $50,000 for his invention in 1794 of the cotton gin, which was easy to imitate. Our own age awarded the Nobel Prize for physics to inventor Jack Kilby in 2000 for the research that led to the integrated circuit.

Yet it should be recognized that contributions to knowledge require more complex incentives than mere property rights. No academic economist should be surprised by a statement that the payoff to successful research is more than just a *financial* compensation or rewards correlated with it. While the twentieth century was the age of profit-driven corporate research, it also witnessed unprecedented flourishing of open-source activity, in which participants were incentivized in ways that transcended simple profit-maximizing behaviour. This is not only the case in certain software-writing enterprises such as Linux or Mozilla. It holds for much of the university-or government-driven programme of scientific research, in which academic researchers increased the body of propositional knowledge by a huge multiple while rarely getting rich in the process. The centrality of university research, supported by government grants in the emergence of many of the major technologies of the late twentieth century, should serve as evidence of the complexities of the motivations of those who add to the stock of propositional knowledge. In research, the name of the game is credit, not profit. Researchers want property rights to their work, but normally prefer peer recognition to a cheque. The results of their work, through complex interactions with technology, have been the cheapest lunch in human history.

Technology and institutions co-evolve, but they obviously follow very different evolutionary dynamics and selection processes; it cannot be expected that their joint evolution will ever result in an optimal environment for technological progress (Nelson 1994). At the same time, the corporate and government sectors have emerged as key players in the creation of new technology. The government sets most of the rules of the game (patent laws and enforcement, antitrust, licensing) as well as some priorities (for example, military and space research, national institutes for health), while corporations in an oligopolistic market maximize profits subject to the institutional structure, and rely on the epistemic bases created mostly by people at universities or research institutions. The net result is imperfect on many levels (especially the now quite problematic patent system). And yet it has produced and will continue to produce a dynamic, innovative society in which technological progress and economic growth have become the rule rather than the exception (Baumol 2002). From a long-term historical point of view, that is quite a miracle.

See Also

- ▶ Growth and Institutions
- ▶ Patents

Bibliography

Aghion, P., and P. Howitt. 1997. *Endogenous growth theory.* Cambridge, MA: MIT Press.

Arora, A., A. Fosfuri, and A. Gambardella. 2004. *Markets for technology: The economics of innovation and corporate strategy.* Cambridge, MA: MIT Press.

Baumol, W. 2002. *The free-market innovation machine*. Princeton: Princeton University Press.

Bresnahan, T., and M. Trajtenberg. 1995. General purpose technologies: Engines of growth? *Journal of Econometrics* 65: 83–108.

Cowan, R., and D. Foray. 1997. The economics of codification and the diffusion of knowledge. *Industrial and Corporate Change* 6: 595–622.

Jaffe, A., and J. Lerner. 2004. *Innovation and its discontents*. Princeton: Princeton University Press.

Khan, B., and K. Sokoloff. 2001. The early development of intellectual property institutions in the United States. *Journal of Economic Perspectives* 15(2): 1–15.

Mansfield, E., J. Rapoport, A. Romeo, S. Wagner, and G. Beardsley. 1977. Social and private rates of return from industrial innovations. *Quarterly Journal of Economics* 91: 221–40.

Mokyr, J. 1990. *The lever of riches: Technological creativity and economic progress*. New York: Oxford University Press.

Mokyr, J. 2002. *The gifts of Athena: Historical origins of the knowledge economy.* Princeton: Princeton University Press.

Nelson, R. 1994. Economic growth through the co-evolution of technology and institutions. In *Evolutionary economics and chaos theory: New directions in technology studies*, ed. L. Leydesdorff and P. Van Den Besselaar. New York: St Martins Press.

Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Belknap Press/ Harvard University Press.

Olsson, O. 2000. Knowledge as a set in idea space: An epistemological view on growth. *Journal of Economic Growth* 5: 253–76.

Polanyi, M. 1962. *Personal knowledge: Towards a post-critical philosophy.* Chicago: Chicago University Press.

Price, D. 1984. Notes towards a philosophy of the science/technology interaction. In *The nature of knowledge: Are models of scientific change relevant?* ed. R. Laudan. Dordrecht: Kluwer.

Robinson, J. 1956. *The accumulation of capital*. Homewood: Irwin.

Rosenberg, N. 1994. *Exploring the black box*. New York: Cambridge University Press.

Rosenberg, N. 1996. Uncertainty and technological change. In *Technology and growth*, Conference series, vol. 40, ed. J. Fuhrer and J. Sneddon Little. Boston: Federal Reserve Bank of Boston.

Smith, M., and L. Marx (eds.). 1994. *Does technology drive history?* Cambridge, MA: MIT Press.

Weitzman, M. 1996. Hybridizing growth theory. *American Economic Review* 86: 207–13.

Temporary Equilibrium

J.-M. Grandmont

Abstract

Temporary general equilibrium views the dynamic evolution of an economy as taking place sequentially in calendar time, with decisions being made and equilibrium being achieved at each date in the light of the traders' expectations about the future. The article

surveys the contributions of the field to the microeconomic foundations of macroeconomics, in particular to the analysis of monetary phenomena, non-clearing markets, imperfect competition and the foundations of Keynesian unemployment, as well as the study of economic dynamics and (in) stability of self-fulfilling expectations under various learning schemes, in relation in particular with 'excess volatility' of financial markets.

Keywords

Arbitrage; Capital market imperfections; Classical dichotomy; Equilibrium theory; Error learning models; E-stability; Excess volatility; Expectations; Fiat money; Futures markets; Game theory; Imperfect competition; Intertemporal equilibrium; Intertemporal substitution effects; Keynesian unemployment; Liquidity trap; Local stability; Microfoundations; Monopolistic competition; New Keynesian macroeconomics; Non-clearing markets; Oligopolistic competition; Quantity theory of money; Real balance effect; Self-fulfilling expectations; Speculation; Staggered price and wage setting; Sticky information; Say's Law; Temporary equilibrium; Uncertainty principle; Walras's Law; Wealth effect

JEL Classifications

D9

The Conceptual Framework

The fact that trade and markets take place sequentially over time in actual economies is a trivial observation. It has nevertheless far-reaching implications. At any moment, economic units have to make decisions that call for immediate action, in the face of a future that is as yet unknown. Expectations about the unknown future play therefore an essential role in the determination of current economic variables. On the other hand, the expectations that traders hold at any time are determined by the information that they have at that date on the economy, in particular on its current and past states. Observed economic processes are thus the result of a strong and complex interaction between expectations of the traders involved and the actual realizations of economic variables.

Economists have long recognized that such an interaction should be at the heart of any satisfactory theory of economic dynamics. The temporary equilibrium approach was indeed designed quite a while ago by the Swedish school (Lindahl 1939) and J.R. Hicks (1939, 1965), with the intent to establish a general conceptual framework that would enable economists to cope with the study of dynamical economic systems, and in particular to incorporate in their models the subtle interplay between expectations and actual realizations of economic variables that seems factually so important. Economic theorists have employed this framework in a systematic way since then, using in particular the powerful techniques of modern equilibrium and/or game theory; this effort has yielded important improvements of our understanding of the microeconomic foundations of macroeconomics.

Before reviewing briefly a few of these important advances, it may be worthwhile to make clear what the basic characteristics of the temporary equilibrium approach are, and to compare it with others. To fix ideas, let us assume that time is divided into an infinite, discrete sequence of dates. We may envision first a specific institutional set-up, that was called a *futures economy* by Hicks (1939), and later generalized by Arrow and Debreu. Let us assume that markets for exchanging commodities are opened at a single date, say date 0; assume further that at that date, markets exist for contracts to deliver commodities at each and every future date $t \geq 0$. The specification of a 'commodity' will then involve not only the physical characteristics of the good or service to be delivered, but also the location and the circumstances ('state of nature') of the delivery. One gets then what has been called a 'complete' set of futures markets at the initial date $t = 0$ (Debreu 1959, ch. 7).

It is clear that this framework is essentially timeless. Once an equilibrium is reached at date

0 (this equilibrium may be Walrasian or the result of any other game theoretic equilibrium notion), production and trade do take place sequentially in calendar time. But the coordination of the decisions of all traders is achieved at a single date through futures markets. There is no sequence of markets over time, and no role for expectations, money, financial assets, or stock markets.

Let us consider next another, more dynamic, type of organization, in which markets do open in every period. In this framework, traders would exchange at every date commodities immediately available on spot markets, promises to deliver specific commodities at later dates on futures markets, as well as money, financial assets and/or stocks (of course markets must be 'incomplete' in the sense of Arrow–Debreu at every date, otherwise reopening markets would serve no purpose). To convey the following discussion most simply, let us assume away all sources of uncertainty and consider the case where the state of the economy at any date can be described by a single real number. To simplify matters further, let us assume that the state of the economy at t, say x_t, is completely determined by the forecasts $x^e_{i,t+1}$ made by all traders $i = 1, \ldots, m$ at date t about the future state, through the relation

$$x_t = f\left(x^e_{1,t+1}, \ldots, x^e_{i,t+1}, \ldots, x^e_{m,t+1}\right) \quad (1)$$

The temporary equilibrium map f describes the result of the market equilibrating process at date t – be it Walrasian or not – for a given set of forecasts. Of course, in the study of any particular economy, the map f will be derived from the 'fundamental' characteristics of the economy: tastes, endowments, technologies, the rules of the game, the policies followed by the government.

The foregoing formulation does seem to take into account the observed fact that markets unfold sequentially in calendar time. It is, however, incomplete since no specification of the way in which forecasts are made at each date has been offered at this stage.

We must first discuss a concept that was introduced by Hicks himself, that of an *intertemporal equilibrium,* with self-fulfilling expectations, and that has been extensively used recently in a variety of contexts. Such an intertemporal equilibrium is defined formally, in the present framework, as an infinite sequence of states $\{x_t\}$ and of forecasts $\left\{x^e_{i,t+1}\right\}$ satisfying (1) and

$$x^e_{i,t+1} = x_{t+1} \quad (2)$$

for all dates. Although time appears explicitly in this formulation, it should be clear that this particular equilibrium concept is also intrinsically *timeless.* Indeed all elements of the sequences of equilibrium states $\{x_t\}$ and of equilibrium forecasts $\left\{x^e_{i,t+1}\right\}$ are determined simultaneously by an outside observer: present and future markets are equilibrated all at the same time.

The preceding discussion shows how we must proceed to describe a sequential adjustment of markets, in calendar time. We *must* add to the temporary equilibrium relationship (1) a specification of the way in which traders forecast the future at each date *as a function of their information on current and past states of the economy.* If we assume, for the simplicity of the exposition, that the information available to traders at date t is represented by the sequence $(x_t, x_{t-1}, \ldots)$, that means that we have to add to (1), m *expectations* functions of the form

$$x^e_{i,t+1} = \psi_i(x_t, x_{t-1}, \ldots) \quad (3)$$

The equations (1) and (3) describe then in a consistent way a sequential adjustment of markets – a sequence of *temporary equilibria* – in which time goes forward, as it should. Given past history $(x_{t-1}, x_{t-2}, \ldots)$, (1) and (3) determine the current temporary equilibrium state and forecasts. Once such a temporary equilibrium is reached, production and exchange takes place at date t, and the economy can move forward to date $t + 1$, where the equilibrating process is repeated.

The temporary equilibrium approach, as sketched in the formulation (1) plus (3) is the general formulation, in fact the *only* sort of formulation that is allowed, if one wishes to describe the evolution of the economy as a *sequence* of

markets that adjust one after each other. One should expect accordingly the approach to include self-fulfilling expectations as a special case. Indeed choose a particular intertemporal equilibrium. Then the associated sequence of states, say $\{\bar{x}_t\}$, is a solution of the difference equation

$$\bar{x}_t = F(\bar{x}_{t+1}) \tag{4}$$

in which $F(x) = f(x, \ldots, x)$ for all x. Consider now the economy at date t, and assume that past states have been $(\bar{x}_{t-1}, \bar{x}_{t-2}, \ldots)$. Assume that the traders know the characteristics of the economy, or at least the map F, and further that the map F is invertible (we are voluntarily vague about the domain of definitions of the functions under consideration, to simplify the present methodological discussion, but these technical details can be fixed up). The traders are then able to infer that the recurrence satisfied by current and past states, that is, $x_t = F^{-1}(x_{t-1})$, will obtain in the future as well, their forecasting rule may be viewed as the result of iterating twice that relation, or of inverting $\bar{x}_{t+1} = F^2(\bar{x}_{t-1})$, for all $i = 1, \ldots, m$

$$\psi_i(\bar{x}_t, \bar{x}_{t-1}, \ldots) = F^{-2}(\bar{x}_{t-1}). \tag{5}$$

If this relation holds, $\bar{x}_t$ is indeed a temporary equilibrium state (that is, it solves (1) and (3)) at date t, given past history $(\bar{x}_{t-1}, \bar{x}_{t-2}, \ldots)$.

As we have just shown, the temporary equilibrium method includes selffulfilling expectations as a special case. This shows incidentally that the opposition often made in the literature, between self-fulfilling expectations, that are claimed to be 'forward looking', and 'backward looking' expectations as in (3), is presumably misleading. The temporary equilibrium approach is indeed much more general, since it permits to incorporate in the analysis the fact that traders usually learn the dynamics laws of their environment only gradually, and thus to study in principle how convergence toward self-fulfilling expectations may or may not obtain in the long run.

The preceding discussion was carried out in a simple one-dimensional world operating under certainty. It should be clear nevertheless that the qualitative conclusions we obtained hold as well in a more complex, multidimensional world operating under uncertainty.

When the evolution of the economy is described as a sequence of temporary equilibria, at each date, the current equilibrium states are determined by past history. In this framework, a number of issues arise naturally. First, one has to find the conditions under which the dynamic evolution of the economy is well defined. In other words, when does a temporary equilibrium exist? Second, does the corresponding dynamical system have long-run equilibrium states, such as deterministic stationary states or cycles, and/or stationary stochastic processes, along which expectations are self-fulfilling? Under which conditions, in particular on the formation of expectations, do the sequences of temporary equilibria so generated converge to such a long-run equilibrium? This is precisely the sort of questions that have attracted the attention of modern economic theorists working in temporary equilibrium theory.

Overview

We turn now to a brief appraisal of this research effort, referring the interested reader to more extensive and more technical surveys that already exist in the literature, see for example Grandmont (1977, 1987, 1998).

Money and Assets in Competitive Markets

Considering a sequence of markets opens immediately the possibility for traders to hold money and more generally, assets of various kinds for saving, borrowing, transactions purposes and/or insurance motives. The application of the modern techniques of temporary equilibrium theory to the study of monetary phenomena has led to a major reappraisal, in the 1970s, of classical and neoclassical monetary theories in competitive environments. It has permitted in particular to solve an old problem that had puzzled economic theorists for some time (Hahn 1965), namely, why fiat money, which has no intrinsic value, should have a positive value in exchange in competitive markets. The answer provided by traditional neoclassical theory relied essentially upon unit-elastic

price expectations and the presence of real balance or wealth effects (Patinkin 1965). Modern temporary equilibrium methods have shown that sort of answer to be surely incomplete and presumably mistaken: intertemporal substitution effects have to play an important role, and this can be achieved only by abandoning the hypothesis of unit-elastic expectations and by introducing some degree of inelasticity of expectations with respect to current observations (an example of such a condition was used in (5) above, where the forecast was made to depend on past states but not on the current state). The reappraisal of monetary theory by means of the temporary equilibrium method clarified greatly many confusing debates of the preceding literature: the relations between Walras's and Say's Law, the meaning and the validity of the classical dichotomy and the quantity theory of money, the possibility of monetary authorities to manipulate the interest rates or the money supply, the existence of a 'liquidity trap' (Grandmont 1983). The introduction of cash-in-advance constraints in temporary competitive equilibrium models of money (Grandmont and Younès 1972, 1973) yielded important insights into the relations between its respective roles as a store of value and as a medium of exchange, and time preference, and permitted to make precise the microeconomic foundations of Milton Friedman's theory of optimum cash balances (1969; see also Woodford 1990). Such models of money using cash-in-advance constraints have been popular in modern macroeconomics, following the contribution of R.E. Lucas, Jr. (1980). Capital market imperfections that make explicit the essential role of money in exchange have since been central to the modern analysis of monetary theory and policy (see Wallace 2001).

The introduction of assets of various kinds in competitive markets leads also to the possibility of speculation and arbitrage in capital markets. Different persons with different tastes or expectations will then be willing to trade such assets. An important question is to study the conditions ensuring the existence of a temporary equilibrium in that context. A neat answer to that problem was provided by J.R. Green (1973) and O.D. Hart (1974): there must be some agreement between the traders' expectations about future prices.

Temporary Equilibria with Non-clearing Markets and Imperfect Competition

A temporary equilibrium need not be Walrasian. One may consider cases where prices and/or wages are set through monopolistic or oligopolistic competition at the beginning of each elementary period and remain temporarily fixed within that period. A temporary equilibrium corresponding to these prices is then achieved at each date by quantity rations that set upper or lower bounds on the traders' transactions.

It had been known for some time that traditional Keynesian macroeconomic models of unemployment involved, explicitly or implicitly, the assumption of temporarily fixed prices and/or wages, as noted by Hicks himself (1965). The choice-theoretic structure of these models was rather unclear, however, which was a source of some confusion. The systematic study of temporary equilibrium models with quantity rationing undertaken in the 1970s produced deep insights on this issue, and unveiled the hidden but central role played by quantity signals, as perceived by the traders in addition to the price system, to achieve an equilibrium in such models.

One major outcome of this research programme was the discovery that different types of unemployment could obtain, and even co-exist. 'Keynesian unemployment' corresponds to a situation where there is an excess supply on the labour and the goods markets. In such a situation, firms perceive constraints on their sales because demand is too low. Keynesian policies aiming at increasing aggregate demand may work in such a case. But unemployment may co-exist with an excess demand on the goods markets. In such a regime, called 'classical unemployment' by Malinvaud (1977), the source of unemployment is rather the low profitability of productive activities. Keynesian policies may not work in that case; one has to resort to policies that restore profits, such as lowering real wages. In that respect, these results achieved a remarkable synthesis, within a unified and clear conceptual

framework, between two paradigms that appeared fundamentally distinct beforehand.

The research on this topic proceeded very early on to endogenize prices and wages and yielded numerous insights of the connections between Keynesian models of unemployment and price or wage making in monopolistic or oligopolistic models of competition (see Barro and Grossman 1976; Benassy 1982, 1986; Grandmont and Laroque 1976; Hart 1982; Malinvaud 1977; Negishi 1979). It has since become a cornerstone building block of the modern reformulation of so-called 'new Keynesian macroeconomics' (Benassy 2002; Dreze 1991; Mankiw and Romer 1991). While early formulations focused on temporary equilibria with non-clearing markets due to exogenously staggered price and wage setting (see Taylor 1999, for a survey), more recent research seeks to explain such staggering of price and wage changes as the rational reaction of agents under the gradual diffusion of 'sticky information' (Mankiw and Reiss 2003).

Learning and (In)stability

The temporary equilibrium approach includes self-fulfilling expectations as a particular case, and is in fact more general, since it can incorporate learning in the formation of the traders' expectations. An important issue, that has been early on the agenda of that research programme (Fuchs and Laroque 1976), is then to know whether the sequences of temporary equilibria that are associated to given learning processes or expectations functions converge eventually to a long-run equilibrium along which forecasting mistakes vanish. The question arises of course for long-run equilibria that are simple, such as steady states, or more complex, such as deterministic cycles (Grandmont 1985). The general lesson that seems to come out the research works done on the topic appears to be some kind of *'uncertainty principle'* (Grandmont 1998). Learning generates local instability of self-fulfilling expectations whenever agents are on average uncertain about the local stability of the system, and thus ready to extrapolate a wide range of regularities (trends) out of past deviations from equilibrium, and when the influence of expectations on the dynamics is significant. On the other hand, learning may generate locally stable dynamics when either expectations do not matter much or traders extrapolate a restricted range of stable trends out of past deviations from equilibrium.

The above principle arises in a wide variety of learning processes, in particular in 'error learning' models, least squares and Bayesian learning. Of course, if one is willing to restrict the range of learning schemes, one may be able to produce sharper stability criteria (see in particular the concept of 'E-stability' developed by Evans and Honkapohja 1999, for a particular class of learning processes). Local learning instability due to the above 'uncertainty principle' may explain why markets in which expectations are thought to play a significant role, such as markets for financial assets, durable goods, capital or inventories, display more volatility than others. In a similar vein, Brock and Hommes (1997) have shown local instability in a cobweb model, and convergence to more or less complex cyclical or chaotic long-run equilibria, when agents choose in variable proportions among different more or less efficient (and costly) learning schemes, and have sought to apply this approach to explain 'excess volatility' in financial markets.

See Also

- ▶ Fiat Money
- ▶ Fixprice Models
- ▶ General Equilibrium
- ▶ New Keynesian Macroeconomics
- ▶ Non-clearing Markets in General Equilibrium
- ▶ Sticky Wages and Staggered Wage Setting

Bibliography

Barro, R.J., and H.I. Grossman. 1976. *Money, employment and inflation*. Cambridge: Cambridge University Press.

Benassy, J.P. 1982. *The economics of market disequilibrium*. New York: Academic.

Benassy, J.P. 1986. *Macroeconomics: An introduction to the non-Walrasian approach*. New York: Academic.

Benassy, J.P. 2002. *The Macroeconomics of imperfect competition and nonclearing markets : A dynamic general equilibrium approach*. Cambridge, MA: MIT Press.

Brock, W.A., and C. Hommes. 1997. A rational route to randomness. *Econometrica* 65: 1059–1095.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium,* Cowles foundation monograph, vol. 17. New York: Wiley.

Dreze, J. 1991. *Underemployment equilibria*. Cambridge: Cambridge University Press.

Evans, G.W., and S. Honkapohja. 1999. Learning dynamics. In *Handbook of macroeconomics*, ed. J.-B. Taylor and M. Woodford, vol. 1. New York: North-Holland.

Friedman, M. 1969. *The optimum quantity of money and other essays*. Chicago: Aldine.

Fuchs, G., and G. Laroque. 1976. Dynamics of temporary equilibria and expectations. *Econometrica* 44: 1157–1178.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45: 535–572.

Grandmont, J.M. 1983. *Money and value: A reconsideration of classical and neoclassical monetary theories,* Econometric society monograph, vol. 5.Cambridge: Cambridge University Press.

Grandmont, J.M. 1985. On endogenous competitive business cycles. *Econometrica* 53: 995–1045.

Grandmont, J.M., ed. 1987. *Temporary equilibrium: Selected readings*. New York: Academic.

Grandmont, J.M. 1998. Expectations formation and stability of large socioeconomic systems. *Econometrica* 66: 741–781.

Grandmont, J.M., and G. Laroque. 1976. On temporary Keynesian equilibria. *Review of Economic Studies* 43: 53–67.

Grandmont, J.M., and Y. Younès. 1972. On the role of money and the existence of a monetary equilibrium. *Review of Economic Studies* 39: 355–372.

Grandmont, J.M., and Y. Younès. 1973. On the efficiency of a monetary equilibrium. *Review of Economic Studies* 40: 149–165.

Green, J.R. 1973. Temporary general equilibrium in a sequential trading model with spot and future transactions. *Econometrica* 41: 1103–1123.

Hahn, F.H. 1965. On some problems of proving the existence of an equilibrium in a monetary economy. In *The theory of interest rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Hart, O.D. 1974. On the existence of equilibrium in a securities model. *Journal of Economic Theory* 9: 293–311.

Hart, O.D. 1982. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97: 109–138.

Hicks, J.R. 1939. *Value and capital*, 2nd ed. Oxford: Clarendon Press, 1946.

Hicks, J.R. 1965. *Capital and growth*. Oxford: Clarendon Press.

Lindahl, E. 1939. *Theory of money and capital*. London: Allen and Unwin.

Lucas, R.E. Jr. 1980. Equilibrium in a pure currency economy. *Economic Inquiry* 18: 203–220. Also in *Models of monetary economies,* ed. J.H. Kareken and N. Wallace, Minneapolis: Federal Reserve Bank of Minneapolis, 1980.

Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Basil Blackwell.

Mankiw, N.G. and Romer, D., eds. 1991. *New Keynesian economics,* 2 vols. Cambridge, MA: MIT Press.

Mankiw, N.G., and R. Reiss. 2003. Sticky information: A model of monetary nonneutrality and structural slumps. In *Knowledge, information and expectations in modern macroeconomics*, ed. P. Aghion et al. Princeton: Princeton University Press.

Negishi, T. 1979. *Microeconomic foundations of Keynesian macroeconomics*. Amsterdam: North-Holland.

Patinkin, D. 1965. *Money, interest and prices,* 2nd ed. New York: Harper & Row.

Taylor, J.B. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of Macroeconomics*, ed. J.B. Taylor and M. Woodford, vol. 1. New York: North-Holland.

Wallace, N. 2001. Whither monetary economics? *International Economic Review* 42: 847–869.

Woodford, M. 1990. The optimum quantity of money. In *Handbook of monetary economics*, ed. B.M. Friedman and F.H. Hahn, vol. 2. New York: North-Holland.

Term Structure of Interest Rates

Burton G. Malkiel

Abstract

The term structure of interest rates concerns the relationship among the yields of bonds that differ only with respect to their terms of maturity. This article explains the three traditional explanations of the term structure. (1) The expectations theory considers the long rate to be an average of current and future short rates. (2) The liquidity-preference theory posits that illiquid, risky long-terms bonds must yield a premium over expected short rates. (3) The hedging-pressure theory stresses the influence of the preferred habitats of different investors. A survey of empirical work on the term structure including affine yield models concludes.

Keywords

Affine models of the term structure; Bonds; Central banks; Expectations theory; Expectations-forming mechanisms; Fisher, I.; Greenspan, A.; Hedging-pressure theory; Hicks, J.; Inflation; Interest rates; Liquidity premium; Liquidity-preference theory; Long-term interest rates; Lutz, F.; No-arbitrage condition; Preferred habitat theory; Risk premium; Short-term interest rates; Term structure of interest rates; Yield curve

JEL Classifications

E43

The term structure of interest rates plays a critical role in the decisions of individuals and corporations and in the conduct of monetary policy. Individuals deciding between an adjustable and a fixed-rate mortgage, and corporations deciding whether to finance their operations with short- or long-term debt, can make sensible decisions only if they understand the factors that determine the relationship between short- and long-term rates. Central banks, which are considered to have substantial control over short-term interest rates, need to understand the likely effect on long rates from their activities in the short-term market.

During 2005, in his final year as Chairman of the Federal Reserve, Alan Greenspan referred to the behaviour of long-term rates as a 'conundrum' (Greenspan 2005, p. 8). The US central bank had raised short-term rates at eleven consecutive meetings but the long-term government-bond rate actually declined over the period. This article examines the Greenspan conundrum and presents the traditional theories offered by economists to explain the relationship between short- and long-term interest rates.

Consider a one-year zero coupon bond that makes a single known payment at maturity, which is F, the face value of the bond. The bond's price is P and the investor's return is R, 1, which is the bond's simple (one-year) yield to maturity:

$$P = \frac{F}{(1+R,1)}.$$

Assuming annual compounding, the yield to maturity of an N-year bond, we can find R, N by solving the equation

$$P = \frac{F}{(1+R,N)^N}.$$

The term structure of interest rates concerns the relationship among the yields of (zero coupon) default-free securities that differ only with respect to their term to maturity. The relationship is more popularly known as the shape of the yield curve, which is pictured by plotting the various yields to maturity (Rs) on the vertical axis against the different years to maturity (Ns) on the horizontal axis. Explaining the shape of curve has been a subject of intense examination by economists for over 60 years. Historically, three competing theories have attracted the widest attention. These are known as the expectations, liquidity preference, and hedging-pressure (or preferred habitat) theories of the term structure.

The Expectations Theory

According to the expectations theory, the shape of the yield curve can be explained by investors' expectations about future short-term interest rates. I use the term 'interest rate' to refer to the yield to maturity of a zero coupon bond of a specific term to maturity (N). This proposition dates back at least to Irving Fisher (1896), but the main development of the theory was done by Hicks (1939) and Lutz (1940). More recent versions of the theory have been developed by Malkiel (1966), Roll (1970, 1971), and Cox et al. (1981).

Suppose, for example, that investors believe that the prevailing level of interest rates is unsustainably high and that lower rates are more probable than higher ones in the future. Under such circumstances, long-term bonds will appear to investors as more attractive than shorter-term

issues *if* both sell at equal yields. Long-term bonds will permit an investor to earn what is believed to be an unusually high rate over a longer period of time than short-term issues, whereas investors in shorter bonds subject themselves to the prospect of having to reinvest their funds later at the lower yields that are expected. Moreover, longer-term bonds are likely to appreciate in value if expectations of falling rates prove correct. Thus, if short and long securities sold at equal yields, investors and arbitrageurs would tend to bid up the prices (force down the yields) of long-term bonds while selling off short-term securities, causing their prices to fall (yields to rise). Thus, a descending yield curve with short issues yielding more than longer ones can be explained by expectations of lower future rates. Similarly, an ascending yield curve, with longer issues yielding more than shorter-term ones, can be explained by expectations of rising rates.

Under the assumptions of the perfect-certainty variant of the expectations theory, there are no transactions costs, and all investors make identical and accurate forecasts of future interest rates. The theory then implies a formal relationship between long- and short-term rates of interest. Specifically, the analysis leads to the conclusion that the long rate is an average of current and expected short rates. Consider the following simple two-period example, where only two securities exist (a one-year and a two-year zero coupon bond), and investors have funds at their disposal for one or two years. Let capital *Rs* stand for actual market rates (yields), while lower-case *rs* stand for expected or forward rates. Prescripts represent the time periods for which the rates are applicable, while postscripts stand for the maturity of the bonds. Thus, t, *R*, 2 indicates today's actual two-year rate, while $t + 1$, r, 1 stands for the expected one-year rate in period $t + 1$.

If investors are profit maximizers, it follows that each investor will choose that security (or combination of securities) that maximizes his return for the period during which his funds are available. Consider the alternatives open to the investor who has funds available for two years. The two-year investor will have no incentive to move from one bond to another when he can make the same investment return from buying a combination of short issues or holding one long issue to maturity. If such an investor invests one dollar in a one-year security and then reinvests the proceeds at maturity (that is, $1 + t$, R, 1) in a one-year issue next year, his total capital will grow to $(1 + t, \mathrm{R}, 1)$ $(1 + t + 1, \mathrm{r}, 1)$ at the end of the two-year period. Alternatively, if he invests his dollar in a two-year zero coupon issue, he will have at maturity $(1 + t, \mathrm{R}, 2)^2$. In equilibrium, where the investor has no incentive to switch from security to security, the two alternatives must offer the same overall yield, namely,

$$(1 + t, R, 2)^2 = (1 + t, R, 1)\ (1 + t + 1, r, 1). \tag{1}$$

Thus, the two-year rate can be expressed as a geometric average involving today's one-year rate and the one-year rate of interest anticipated next year:

$$(1 + t, R, 2) = [(1 + t, R, 1)(1 + t + 1, r, 1)]^{1/2}. \tag{2}$$

If Eq. 2 holds, then the holding-period return for the one-year investor will also be the same whether he buys a one-year bond and holds it to maturity or buys a two-year bond and sells it after one year. If Eq. 2 does not hold, say because the two-year rate was lower than the average of the current and prospective one-year rate, an arbitrageur could make a sure profit by selling the two-year issue short and purchasing a series of one-year securities. It is in this sense that the expectations theory fulfills a no-arbitrage condition.

In similar fashion, the rate on longer-term issues must turn out to be an average of the current and a whole series of future short-term rates of interest. Only when this is true can the pattern of short and long rates in the market be sustained. The long-term investor must expect to earn through successive investment in short-term securities the same return over his investment period that he would earn by holding a long-term bond to maturity. In general, the equilibrium relationship is,

$$(1+t,R,N) = [(1+t,R,1)(1+t+1,r,1) \ldots\ldots (1+t+N-1,r,1)]^{1/N} \quad (3)$$

The expectations theory can be extended to a world of uncertainty and it can account for every sort of yield curve. If short-term rates are expected to be lower in the future, then the long rate, which we have seen must be an average of those rates and the current short rate, will lie below the short rate. Similarly, long rates will exceed the current short rate if rates are expected to be higher in the future.

Notice that the expectations theory is capable of explaining the Greenspan conundrum. Suppose that the central bank's aggressive raising of short-term rates was expected to lower future inflation and weaken future economic activity, thus leading to expectations of lower future short-term rates. In such a case, the long rate could fall because, while the current short rate was higher, the entire set of future short rates would be lower than was previously expected.

The Liquidity-Preference Theory

The liquidity-preference theory, advanced by Hicks (1939), accepts that expectations are important in influencing the shape of the yield curve. Nevertheless, it argues that, in a world of uncertainty, short-term issues are more desirable to investors than longer-term issues because they are more liquid. Short-term issues can be converted into cash at short notice without appreciable loss in principal value, even if rates change unexpectedly. Long-term issues, however, will tend to fluctuate widely in price with unanticipated changes in interest rates and hence ought to yield more than shorts by the amount of a risk premium.

If no premium were offered for holding long-term bonds, it is argued that most individuals and institutions would prefer to hold short-term issues to minimize the variability of the money value of their portfolios. On the borrowing side, however, there is assumed to be an opposite propensity. Borrowers can be expected to prefer to borrow at long term to assure themselves of a steady source of funds. This leaves an imbalance in the pattern of supply and demand for the different maturities – one that speculators might be expected to offset. Hence, the final step in the argument is the assertion that speculators are also averse to risk and must be paid a liquidity premium to induce them to hold long-term securities. The arbitrage described in the exposition of the expectations theory is not riskless. Thus, even if interest rates are expected to remain unchanged, the yield curve should be upward sloping, since the yields of long-term bonds will be augmented by risk premiums necessary to induce investors to hold them. While it is conceivable that short rates could exceed long rates, if investors think that rates will fall sharply in the future, the 'normal relationship' is assumed to be an ascending yield curve.

Formally, the liquidity premium is typically expressed as an amount that is to be added to the expected future rate in arriving at the equilibrium-yield relationships described in Eqs. 1 through 3. If we let L, 2 stand for the liquidity premium that should be added to next year's forecasted one-year rate, we have

$$(1+t,R,2)^2 = (1+t,R,1) \times (1+t+1,r,1+L,2) \quad (4)$$

and

$$(1+t,R,2) = [(1+t,R,1)(1+t+1,r,1+L,2)]^{1/2}. \quad (5)$$

Thus, if L, 2 is positive (that is, if there is a liquidity premium), the two-year rate will be greater than the one-year rate even when no change in rates is expected. It has also been customary to assume that L, 3, the premium to be added to the one- year rate forecast for two years hence (that is, period $t+2$), is even greater than L, 2, so that the three-year rate will exceed the two-year rate when no change is expected in short-term rates over the next three years. In general, the liquidity-premium model may be written as

$$(1+t,R,N) = [(1+t,R,1)(1+t+1,r,1+L,2) \ldots\ldots (1+t+N-1,r,1+L,N)]^{1/N}. \quad (6)$$

On the assumption that $L, N > L, N-1 > \ldots > L, 2 > 0$, the yield curve will be positively sloped even when no changes in rates are anticipated.

Another potential explanation of the Greenspan conundrum is consistent with the liquidity-preference theory. The prompt actions of the central bank to insure that future inflation is contained could engender expectations of greater future economic stability and thus lower risk premiums throughout the economy.

The Hedging-Pressure or Preferred Habitat Theory

Other critics of the expectations theory, including Culbertson (1957) and Modigliani and Sutch (1966), argue that liquidity considerations are far from the only additional influence on bond investors. While liquidity may be a critical consideration for a commercial banker considering an investment outlet for a temporary influx of deposits, it is not important for a life insurance company seeking to invest an influx of funds from the sale of long-term annuity contracts. Indeed, if the life insurance company wants to hedge against the risk of interest-rate fluctuations, it will prefer long, rather than short, maturities. Long-term investments will guarantee the insurance company a profit regardless of what happens to interest rates over the life of the contract.

Many pension funds and retirement savers find themselves in a wholly analogous situation. A retirement saver who has funds to invest in bonds for n periods will find an n-period pure discount (zero coupon) bond to be the safest investment. It is assumed that, if investors are risk averse, they can be tempted out of their preferred habitats only with the promise of a higher yield on a bond of any other maturity. Of course, other investors such as commercial banks or corporate investors will hedge against risk by confining their purchases to short-term issues. These investors will need higher yields on longer-term issues to induce them to invest in such securities. Under this hedging-pressure theory, however, there is no reason for term premiums to be necessarily positive or to be an increasing function of maturity. Under an extreme (and somewhat implausible) form of the argument suggested by Culbertson, the short and long markets are effectively segmented, and short and long yields are determined by supply and demand in each of the segmented markets.

One of the popular explanations of the Greenspan conundrum was that many long-term investors, such as pension funds, were moving money out of the stock market and into the long-term bond market in order to 'immunize' their long-term liabilities. Such buying creates additional demand for long-term bonds, driving the yields of such securities down.

Empirical Analysis of Term Structure Theories

The chief obstacle to effective empirical analysis of the determinants of the term structure of interest rates has been the lack of independent evidence concerning expectations of future interest rates. Consequently, the first step in most empirical tests of the pure form of the expectations theory has been to set up some mechanism by which expectations may reasonably have been formed by market participants.

Since people usually estimate the future by relying, at least in part, on historical information, this procedure has often involved the generation of forecasts of future interest rates from past values of these rates. Then investigators have sought to determine whether empirical yield curves have been consistent with these hypothetical forecasts and with the premise that investors, in fact, behave as the expectations theory claims. Thus, in essence, two theories were tested jointly: first, a theory of expectations formation, and second, a theory of the term structure. Of course, it is important to realize that any inability to confirm the expectations theory may be due to a failure

to specify properly an expectations-forming mechanism rather than a failure of the theory to offer a correct explanation of the shape of the yield curve. Nevertheless, the wide body of evidence we have does suggest a general conclusion.

The expectations-forming mechanisms utilized in empirical studies have been varied and inventive. They have included an error-learning mechanism (Meiselman 1962); distributed lags on past rates (Modigliani and Sutch 1966) or on inflation (Modigliani and Shiller 1973; Fama 1976); use of *ex post* data under an assumption that market efficiency and rationality require that *ex post* realizations do not differ systematically from ex ante views (Roll 1970, 1971; Fama 1984a, b); and survey data assumed to reflect the actual expectations of market participants (Kane and Malkiel 1967; Malkiel and Kane 1969; Kane 1983). While affirming the general importance of expectations in influencing the shape of the yield curve, empirical studies have generally rejected the pure form of the expectations hypothesis. There does appear to be an upward bias to the shape of the yield curve, indicating that term premiums do exist. But, contrary to the liquidity-preference theory, term premiums do not increase monotonically over the whole span of forward rates. Moreover, such term premiums vary over time. In addition, there appear to be seasonal patterns in the forward rates calculated from the short end of the yield curve.

Campbell (1995) points out that the pure expectations hypothesis implies that, whenever long yields exceed short yields, short yields should tend to rise in the future. It also implies that long yields must rise in the future so as to produce the capital losses that equate the short-term holding period returns between long and short maturities. He finds, instead, that mean excess returns on long bonds are positive. But excess returns on longer-term bonds do not rise throughout the maturity spectrum. The excess returns on zero coupon two-month bonds over one-month bills are positive, and excess returns over short bills rise with maturity at first, but after one year begin to decline and actually become negative for 10-year zero coupon bonds. And when the long-short yield spread is high, long yields have tended to fall, thus amplifying the yield differential between long and short bonds. Note that it is precisely this behaviour that Chairman Greenspan referred to as a 'conundrum' during 2005.

Affine Yield Models

More recent work on the term structure of interest rates has focused directly on how the shape of the term structure changes over time. The literature has evolved mostly in continuous time, and it is assumed that the future dynamics of the term structure depend on the evolution of some single factor or multiple factors that follow a stochastic process. The models are rooted in a framework consistent with the risk-adjusted expectations theory where arbitrage opportunities are not possible in equilibrium and where (log) bond yields are functions (which are often 'affine') of a single state variable that describes movements in future short-term rates or in a set of state variables related to the workings of the economy and the formation of expectations.

The pioneering models of this type were presented by Vasicek (1977) and Cox et al. (1985). These early papers developed single-factor models where the specific factor was the very short-term (instantaneous) default-free interest rate. Thus, all the information that is relevant for the determination of long-term rates was compressed into one stochastic process for very short rates. The process is a continuous time analogue to an autoregressive process where there is a fixed 'normal' short-term rate that can serve as an anchor for mean reversion. Once the diffusion process for short rates is specified, arbitrage is relied upon to explain the observed yields of bonds of different maturities. The models described above are special cases of the affine class of term structure models.

The models that followed have employed multiple factors but maintained the assumption that (log) bond prices are linear functions of the state variables. Duffie and Kan (1996) and Singleton and Dai (2000) presented models where the

factors are the yields of n various fixed maturity bonds. Litterman and Scheinkman (1999) show that at least 95 per cent of the variation in yield changes can be explained by three latent factors, and interpret these factors in terms of the 'level' of yields, the 'slope' of the yield curve, and the 'curvature' of the curve. Further work has made significant strides in identifying and relaxing the restrictive assumptions of these models and in improving estimation techniques. Dewachter and Lyrio (2006) have jointly modelled the term structure as well as the dynamics of various macroeconomic variables. Their paper provides a macroeconomics interpretation of the Litterman–Scheinkman latent factors. They interpret the 'level' of yields factor as representing inflation expectations, the 'slope' of the yield curve factor as representing business-cycle conditions, and the 'curvature' factor as representing an independent monetary policy factor. Note that these factors are exactly those referenced earlier in an attempt to explain the Greenspan conundrum. A full discussion of affine yield models is contained in a survey article by Piazzesi (2006).

See Also

- ▶ Affine Term Structure Models
- ▶ Bonds
- ▶ Capital Asset Pricing Model
- ▶ Efficient Markets Hypothesis
- ▶ Risk Aversion

Bibliography

Campbell, J. 1995. Some lessons from the yield curve. *Journal of Economic Perspectives* 9 (3): 129–152.

Cox, J., J. Ingersoll, and S. Ross. 1981. The relationship between forward prices and futures prices. *Journal of Financial Economics* 9: 321–346.

Cox, J., J. Ingersoll, and S. Ross. 1985. A theory of the term structure of interest rates. *Econometrica* 53: 385–407.

Culbertson, J. 1957. The term structure for interest rates. *Quarterly Journal of Economics* 71: 485–517.

Dewachter, H., and M. Lyrio. 2006. Macro factors and the term structure of interest rates. *Journal of Money, Credit, and Banking* 38: 119–140.

Duffle, D., and R. Kan. 1996. A yield factor model of interest rates. *Mathematical Finance* 6: 379–406.

Fama, E. 1976. Forward rates as predictors of future spot rates. *Journal of Financial Economics* 3: 361–377.

Fama, E. 1984a. The information in the term structure. *Journal of Financial Economics* 13: 509–528.

Fama, E. 1984b. Term premiums in bond returns. *Journal of Financial Economics* 13: 529–546.

Fisher, I. 1896. Appreciation and interest. *Economic Journal* 6: 567–570.

Greenspan, A. 2005. Statement of Alan Greenspan, Chairman, Board of Governors of the Federal Reserve System, before the Committee on Banking, Housing and Urban Affairs. United States Senate. February 16, 2005. http://banking.senate.gov/_files/ACF5A5C.pdf. Accessed 20 Jan 2006.

Hicks, J. 1939. *Value and capital: An inquiry into some fundamental principles of economic theory*. Oxford: Clarendon Press. 1974.

Kane, E. 1983. Nested tests of alternative term-structure theories. *Review of Economics and Statistics* 65: 115–123.

Kane, E., and B. Malkiel. 1967. The term structure of interest rates: An analysis of a survey of interest-rate expectations. *Review of Economics and Statistics* 49: 343–355.

Litterman, R., and J. Scheinkman. 1999. Common factors affecting bond returns. In *The debt market*, ed. S. Ross. Northampton: Edward Elgar.

Lutz, F. 1940. The structure of interest rates. *Quarterly Journal of Economics* 55: 36–63.

Malkiel, B. 1966. *The term structure of interest rates: Expectations and behavior patterns*. Princeton: Princeton University Press.

Malkiel, B., and E. Kane. 1969. Expectations and interest rates: A cross-sectional test of the error-learning hypothesis. *Journal of Political Economy* 77: 453–470.

Meiselman, D. 1962. *The term structure of interest rates*. Englewood Cliffs: Prentice Hall.

Modigliani, F., and R. Shiller. 1973. Inflation, rational expectations and the term structure of interest rates. *Economica* 40: 12–43.

Modigliani, F., and R. Sutch. 1966. Innovations in interest rate policy. *American Economic Review* 56: 178–197.

Piazzesi, M. 2006. Affine term structure models. In *Handbook of financial econometrics*, ed. Y. Ait-Sahalia and L. Hansen. Amsterdam: North-Holland.

Roll, R. 1970. *The behavior of interest rates: An application of the efficient market model to U.S. treasury bills*. New York: Basic Books.

Roll, R. 1971. Investment diversification and bond maturity. *Journal of Finance* 26: 51–66.

Singleton, K., and Q. Dai. 2000. Specification analysis of affine term structure models. *Journal of Finance* 55: 1943–1978.

Vasicek, O. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5: 177–188.

Terms of Trade

Ronald Findlay

Abstract

This article describes the various approaches that have been made to the 'terms of trade', the question of determining and measuring the rate at which countries that engage in international trade divide the benefits between them. It traces the evolution of the relevant ideas, beginning with the pioneering contributions of David Ricardo, John Stuart Mill and Alfred Marshall.

Keywords

Balance of trade; Barter; Commodity terms of trade; Comparative advantage; Double factoral terms of trade; Gains from trade; Gross barter terms of trade; Immiserizing growth; Income terms of trade; Index numbers; International trade theory; Marshall, A.; Marshall–Lerner condition; Mill, J. S.; North–South economic relations; Offer curve; Optimal tariffs; Prebisch, R.; Reciprocal supply and demand; Ricado, D.; Shadow pricing; Singer, H.; Single factoral terms of trade; Stationary state; Terms of trade; Transfer problem

JEL Classifications

F1

The two most basic questions about international trade are, 'What goods will each country export?' and 'What will be the ratios at which the exports of one country exchange for those of its trading partners?'

The first problem is that of 'comparative advantage'; the second that of the 'terms of trade', which is the subject of the present article. David Ricardo, in Chapter 7 of the Principles (1817), gave a definitive answer to the first question and went a long way towards the solution of the second, though it was John Stuart Mill and Alfred Marshall who eventually gave the complete answer.

In the following discussion it will be convenient to assume, for simplicity, that there is only a single good exported and imported, and sometimes even that there is only a single factor of production, such as labour of a given quality. In practice, of course, we would have to use index numbers for unit values and physical volumes of exports and imports, giving rise to all the familiar problems (see index numbers).

Concepts and Definitions

There are a number of alternative concepts and associated statistical measures of the terms of trade. The most prominent are listed below.

1. The commodity or net barter terms of trade is by far the most common meaning of the term, and is usually what is meant when the expression is used without any qualifying prefix. In principle, this is the relative price of the 'exportable' in terms of the 'importable', the number of units of the latter obtainable for each unit of the former. It has the dimensions of 'nine waistcoat buttons for a copper disc' in the words of Lewis Carroll's 'Song of the Aged, Aged Man' in *Through the Looking Glass,* words that D.H. Robertson (1952, ch. 13) used as the motto for a delightful essay on the terms of trade. In statistical practice, the commodity terms of trade are calculated as changes in the ratio of an export price index to an import price index, relative to a base year.
2. The *gross barter* terms of trade is a concept introduced by F.W. Taussig. It is the ratio of the *volume* of imports to the *volume* of exports. It coincides with the commodity terms of trade when trade is balanced, that is, there are no international loans or unrequited transfers. A deficit in the trade balance would cause the gross barter terms to be more favourable than the commodity or net barter terms, and vice versa. This should not, of course, be understood to mean that a trade deficit is necessarily

preferable to balanced trade, since the additional imports now may have to be paid for by future trade surpluses.

3. The *income* terms of trade, sometimes also referred to as 'the purchasing power of exports' corresponds to the commodity terms of trade multiplied by the volume of exports. This is equal to the volume of imports under balanced trade, and exceeds or falls short of it if there is a surplus or deficit, respectively, in the balance of trade. In other words, it is the level of imports in real terms that can be sustained by current export earnings.
4. The single factoral terms of trade refers to the marginal or average productivity of a factor in the export sector, evaluated in terms of the imported good at the commodity terms of trade. The concept is meaningful for any single factor or production taken separately, though it is sometimes defined in a non-operational fashion in the literature as referring to 'units of productive power'.
5. The double factoral terms of trade is an attempt to go behind the international exchange of commodities to the productive factors that are 'embodied' in them. Thus, if units are chosen, to the effect that a unit of labour in England produces a unit of cloth and a unit of labour in Portugal produces a unit of wine, commodity terms of trade of say five wine to one cloth would mean that a unit of English labour exchanges implicitly for five units of Portuguese labour in international trade.

The first three concepts of the terms of trade are all measurable in practice, subject to the usual index number problems. The commodity terms of trade are routinely calculated for most countries in the world by international agencies such as the United Nations (UN), the World Bank and the International Monetary Fund (IMF). The gross barter and income terms of trade have also been calculated for several countries.

The single factoral terms of trade, for any particular factor, can also be computed. Indeed it corresponds exactly to the concept of 'shadow prices' for primary inputs that has recently been developed in the literature on cost–benefit analysis in distorted open economies. Thus it could indicate what the value of a worker or an acre of land, engaged say in the coffee export sector, was worth in terms of imported food at the commodity terms of trade. This could serve as a valuable guide to resource allocation if a comparison were made with what these resources could produce in the domestic food sector.

The double factoral terms of trade, however, is either misleading if it is computed for any particular single factor in a world of more than one scarce input, or non-operational if defined amorphously as applying to units of 'productive power'. This concept has been regarded by more than one economist as the fundamental one, and no less an authority than D.H. Robertson, in the essay referred to above, called it the 'true' terms of trade. Equally eminent authorities, such as Haberler (1955) and Viner (1937), have, however, been more sceptical, and rightly so.

The concept has recently come to the fore again, after many decades of neglect, in connection with the theories of A. Emmanuel (1972) on 'unequal exchange' in trade between high-wage and low-wage countries, which he regards as a form of 'exploitation' of the latter by the former. It is possible to interpret Emmanuel as saying that it is only when the double factoral terms of trade are equal to unity that there is no unequal exchange. As Emmanuel himself acknowledges, however, his argument requires equal capital intensity in the export sectors of the trading partners. We may all agree with Robert Burns that 'A Man's a Man for a' That' in terms of dignity and spiritual worth. It is another thing, however, to say that skill or physical capital, both accumulated at some cost, should count for nothing, and that the only 'fair' exchange is one that takes place according to the simple labour theory of value.

Furthermore, it is clear that the commodity terms of trade can improve while the factoral terms worsen, and vice versa. Thus suppose that, initially, one day's labour in 'North' and 'South' produces a unit of steel and coffee, respectively, and that the commodity terms of trade was one unit of steel for one unit of coffee. Suppose now that one worker in the North produces three steel, while his counterpart in the South still produces

only one coffee. Let the commodity terms of trade now be two units of steel for one of coffee. The commodity terms have doubled in favour of the South, while its factoral terms have deteriorated to two-thirds instead of unity. Which situation would the South prefer?

Fundamental Determinants

Ricardo did not determine the terms of trade explicitly in his analysis in Chapter 7 of the *Principles*. He was only able to show that the equilibrium value would be between the comparative cost ratios of the two countries, specified by the linear technologies. It was John Stuart Mill (1844) who solved the problem by his numerical example of 'reciprocal supply and demand', later refined by Marshall (1930) through the geometric device of the 'offer curves' showing the excess supplies and demands of the two goods in each country as functions of the terms of trade, the equilibrium value of which would be determined by setting world excess supply equal to zero. Marshall demonstrated the possibility of multiple equilibria and also established a criterion for stability of equilibrium that is in use to this day, in the form of the so-called Marshall-Lerner condition that the sum of the import demand elasticities has to be greater than unity.

In modern terms it is the preferences of the consumers that have to be introduced to close the model. Once these are introduced the equilibrium value(s) of the terms of trade are determined as a function of these preferences, the labour endowments and the technical coefficients of production. The subsequent development of the literature has generalized Ricardo's analysis to any number of goods, factors and countries and to variable instead of fixed technical coefficients.

The determination of the terms of trade is thus technically nothing other than that of finding the equilibrium vector(s) of relative prices for general equilibrium models in which there is a world market for tradable goods and internationally mobile factors, and national markets for non-traded goods and internationally immobile factors.

In addition to constituting a central problem for the theory of international trade in its 'positive' aspect, the terms of trade plays if anything an even more critical role in the 'normative' dimension of evaluating the 'gains from trade'. It is crucial to keep these two facets of the terms of trade conceptually distinct, though of course they are both involved in almost every theoretical or policy problem. Another essential distinction is between the terms of trade as an exogenously determined parameter, as in the 'small' open economy models, and as an endogenously determined variable, the equilibrium value of which is altered by some change in circumstances or parameters, such as factor endowments, technology or tastes. Much confusion has been caused in the literature by failure to bear these basic distinctions in mind at all times.

In the realm of positive theory, the terms of trade generally appear in comparative static exercises as the key dependent variable, upon which the effect of some exogenous shock is sought. As an example, consider the effect of a switch in the composition of home demand in favour of the imported commodity. At constant terms of trade this would create an excess demand for the imported good. On the assumption of Walrasian stability, this must lead to a deterioration of the home country's terms of trade for the world market to return to equilibrium.

The famous 'transfer problem' is another example of this sort of comparative static exercise. The transfer of purchasing power at constant terms of trade would lead to an excess supply in the world market of the transferor's exportable, if the home propensity to consume this good is greater than that of the recipient country (the so-called 'classical presumption'). Thus the terms of trade of the transferor would deteriorate, given Walrasian stability, imposing a 'secondary burden' on the transferor.

Finally, we may consider the effects of economic growth, in the form of exogenous changes in factor endowment or technical innovations in either sector, a literature that was stimulated by Hicks's (1953) inaugural lecture on the 'dollar shortage'. Here again the analysis consists in finding the effect of the change on excess supply or demand at constant terms of trade, and thus

obtaining the direction of movement in the terms of trade necessary to clear the market, assuming stability in the Walrasian sense.

Welfare Effects

All these exercises in positive theory of course have welfare consequences for both trading partners. In the case of the two-country transfer problem the transferor is worse off, even if the terms of trade were to move in its favour, while the recipient is better off, even if the terms of trade were to turn against it. In the case of a shift in the composition of home demand towards imports the welfare of the trading partner will rise under normal conditions (see below) as a result of this improvement in its terms of trade. If growth in one country creates an excess demand for imports at constant terms of trade, its passive partner will also benefit from the resulting increase in the relative price of its export.

In the last two cases a country experiences an exogenous improvement in its terms of trade, with no alteration in its own preferences, technology or factor endowment. Must its welfare necessarily increase as a result? The answer in general is 'yes', unless there are domestic distortions such as monopoly or monopsony in product or factor markets, exogenous wage differentials or real factor–price rigidities. A simple example of how it is possible for a country to experience a loss in welfare as a result of an improvement in its commodity terms of trade can be constructed as follows. Suppose that domestic production is completely specialized on the export good and that the real wage is fixed in terms of the imported good. At constant employment, and therefore constant marginal physical productivity of labour in terms of the exported good, the real wage would be lower in terms of the imported good because of the improvement in the terms of trade. This will induce a decline in employment and output until the marginal physical product of labour rises in the same proportion as the relative price of the imported good has fallen, so that the original level of the real wage is restored. The terms of trade improvement, given employment and output, increases welfare, but the contraction in these variables induced by the change in the terms of trade reduces welfare. This negative effect can clearly be sufficient to outweigh the positive effect of the terms of trade gain considered in isolation, since the counteraction can be very sharp if the marginal productivity of labour schedule is assumed to be sufficiently elastic.

Haberler (1955, p. 30), in a characteristically penetrating and judicious discussion of the subject, has stated that 'other things being equal an improvement in the commodity terms of trade does imply an increase in real national income'. As our analysis of the example in the previous paragraphs shows, however, even such a cautious formulation needs to be interpreted with care. It would obviously be a mistake to compound the welfare effects of an exogenous shift in the terms of trade with the direct welfare effects of some independent shock. In our example, however, the terms of trade change was the sole shift in the data, the contraction of employment and output being induced by this very change in the terms of trade itself.

When the change in the terms of trade is a consequence of some exogenous shock, such as a change in tastes, technology or factor endowment, it is clearly erroneous to infer the total change in welfare solely from the direction of change in the terms of trade. Technological progress in the export sector that causes deterioration in the terms of trade can obviously leave a country better off in spite of the deterioration, even though it would of course have been still better off if the terms of trade had remained unchanged. It is this sort of consideration that has led to the introduction of concepts such as the factoral terms of trade, since these measures could show an improvement even when the commodity terms of trade deteriorate. In general, however, it is a mistake to expect any single concept of the terms of trade to be an unambiguous indicator of changes in the gains from trade when there are shifts in the fundamental determinants of tastes, technology and factor endowments.

The welfare effects of such changes can be broken into two parts: first, the effect at unchanged terms of trade and second, the effect of the associated change in the terms of trade. The

net effect on welfare may thus be positive or negative and need not correspond with the direction of the change in the terms of trade. Bhagwati (1958) established the possibility that the net effect on welfare of the country experiencing economic growth can be negative, a phenomenon that he termed 'immiserizing growth'.

Finally, we may consider the terms of trade as an objective of policy, when the country has some degree of monopoly power in international markets. The consideration of a rational policymaker, ignoring the possibility of retaliation, would be to restrict trade to such an extent as to equate at the margin the benefit resulting from the improvement in the terms of trade with the loss of welfare resulting from the decline in the volume of trade. This is the famous 'optimum tariff' argument, the level of which varies inversely with the elasticity of foreign demand for imports.

Secular Tendencies

In addition to comparative static analyses of the type considered up to now, the literature also contains some more speculative hypotheses about secular tendencies in the terms of trade. In the Ricardian tradition capital accumulation and technical progress lead to a steady expansion in the supply of manufactures, while the supply of primary products is always constrained by the limited availability of 'land' and other natural resources. Ricardo's theorem for a closed economy – that growth would raise the relative price of food and therefore the rent of land until a 'stationary state' is approached – has been extended to the world economy in the form of a presumption that there would be a tendency for the terms of trade to move against manufactures and in favour of primary products. Keynes, Beveridge, Robertson and E.A.G. Robinson all took part in a long-running debate on this issue. The story that W.S. Jevons kept enormous stocks of coal in his basement is a bizarre manifestation of this phobia. W.W. Rostow (1962, chs 8 and 9) gives a very interesting review and analysis of this literature, which foreshadows the views associated with the Club of Rome on the depletion of exhaustible natural resources.

Discussions of the secular tendencies of the terms of trade since the Second World War, however, have been dominated by the view of Raul Prebisch (1950) and Hans Singer (1950) that the historical record shows a long-run tendency for the commodity terms of trade of the less developed countries to deteriorate. The evidence was a series showing an apparent long-run improvement in Britain's terms of trade between 1870 and 1940. Theoretical reasons given for the alleged tendency have been lower income-elasticity of demand for primary products than for manufactures, technical progress that economizes on the use of imported raw materials and monopolistic market structures in the industrial countries combined with competitive conditions in the supply of primary products. The general consensus on the statistical debate that has arisen on this issue is that there has not been any discernible secular trend for the commodity terms of trade of the developing countries to deteriorate (see Spraos 1980, for a summary and assessment of the evidence; Lewis 1969, presents an interesting alternative theoretical and empirical analysis of this problem). Hadass and Williamson (2001) provide a useful recent summary and critique of the ongoing debate on the Prebisch–Singer thesis.

The Prebisch–Singer hypothesis and the more general concerns of the ongoing North–South dialogue have also spawned a number of so-called 'North–South' models, in which the interaction of an advanced industrial region with a less developed and structurally dissimilar, labour-abundant, primary producing region is studied in a dynamic context. The terms of trade play a key role in these models, since the growth rate of the South is linked to this variable through dependence on capital goods imported from the North. This and other analytical issues related to secular trends in the terms of trade are further discussed in Findlay (1980, 1984), Taylor (1981) and Darity (1990).

See Also

► Comparative Advantage
► Heckscher–Ohlin Trade Theory

▶ Index Numbers
▶ Prebisch, Raúl (1901–1986)

Bibliography

Bhagwati, J. 1958. Immiserizing growth: A geometrical note. *Review of Economic Studies* 25: 201–205.
Darity, W. 1990. Fundamental determinants of the terms of trade reconsidered: Long-run and long-period equilibrium. *American Economic Review* 80: 816–827.
Emmanuel, A. 1972. *Unequal exchange*. New York: Monthly Review Press.
Findlay, R. 1980. The terms of trade and equilibrium growth in the world economy. *American Economic Review* 70: 291–299.
Findlay, R. 1984. Growth and development in trade models. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen, vol. 1. Amsterdam: North- Holland.
Haberler, G. 1955. *A survey of international trade theory.* Princeton: International Finance Section.
Hadass, Y., and J.G. Williamson. 2001. *Terms of trade shocks and economic performance 1870–1940: Prebisch and Singer revisited*, Working paper no. 8188. Cambridge, MA: NBER.
Hicks, J.R. 1953. An inaugural lecture. *Oxford Economic Papers* 5: 117–135.
Lewis, W.A. 1969. *Aspects of tropical trade 1883–1965*. Stockholm: Almqvist and Wiksell.
Marshall, A. 1930. *The pure theory of foreign trade.* Clifton: Augustus M. Kelley, 1974. (Privately printed, 1879).
Mill, J.S. 1844. Of the laws of interchange between nations; and the distribution of the gains of commerce among the countries of the commercial world. In *Essays on some unsettled questions of political economy.* London: London School of Economics, 1948.
Prebisch, R. 1950. *The economic development of Latin America and its principal problems*. New York: United Nations.
Ricardo, D. 1817. *On the principles of political economy and taxation*. Cambridge: Cambridge University Press, 1951.
Robertson, D.H. 1952. *The terms of trade: Chapter 13 of utility and all that*. New York: Macmillan.
Rostow, W.W. 1962. *The process of economic growth*. New York: Norton.
Singer, H.W. 1950. The distribution of gains between investing and borrowing countries. *American Economic Review, Papers and Proceedings* 5: 473–485.
Spraos, J. 1980. The statistical debate on the net barter terms of trade between primary commodities and manufactures. *Economic Journal* 90: 107–128.
Taylor, L. 1981. North-south trade and southern growth: Bleak prospects from a structuralist point of view. *Journal of International Economics* 11: 589–601.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper Bros.

Terms of Trade and Economic Development

H. W. Singer

One of the most widely discussed theories concerning the terms of trade of developing countries is the Prebisch–Singer hypothesis, independently published in 1950 (Prebisch 1950; Singer 1950). This hypothesis proclaimed a structural tendency for the terms of trade of developing countries to deteriorate in their dealings with industrial countries. In the original form this related mainly to the terms of trade between primary commodities and manufactured goods from the industrial countries. The historical statistical basis was an analysis of British terms of trade during the period 1873–1938 which corresponded to this image of exports of manufactured goods in exchange for primary commodities.

During the first half of the 19th century the historical statistical experience regarding British terms of trade was in the opposite direction. British import prices of primary commodities such as cotton, wool, etc. increased in relation to the prices of British manufactured products (with textile manufactures prominent among exports at that time). This was in line with classical thinking according to which there would be diminishing returns in the production of primary products, due to the scarcity of land and mineral resources (Malthus, Ricardo and extended by Jevons to the cases of coal and minerals more generally). In classical thinking, up to and including John Stuart Mill, it was taken for granted that there was a tendency for the prices of primary commodities to rise in relation to manufactures, especially since the pressure of surplus population and the process of urbanization would keep wages and cost of production in manufacturing low; this was indeed in line with actual experience in the first half of the 19th century and formed the basis of Marx's theory of surplus value and was later applied by Arthur Lewis to conditions in developing countries in his emphasis on the role of unlimited

supplies of labour in economic development (Lewis 1954).

Thus when Singer in 1947/48 prepared for the United Nations his analysis of British terms of trade after 1873 (*Relative Prices of Exports and Imports of Under-developed Countries*, New York: United Nations, 1949) which subsequently formed the basis of the Prebisch–Singer hypothesis, this ran contrary to traditional thinking. Hence there was a great reluctance even to accept the empirical evidence for this period. In particular the question of transport costs and also the question of improving quality of manufactured goods were used by critical economists to contest the empirical basis of the UN study and the Prebisch–Singer hypothesis (Viner 1953; Haberler 1961; Ellsworth 1956; Morgan 1959). However, subsequent analysis has shown that correction for shipping costs and changing quality would not destroy the empirical basis for the hypothesis (Spraos 1980, 1983).

The extension of the Prebisch–singer hypothesis to the post-war period has also been questioned empirically. At the time the hypothesis was formulated in 1949/50 primary commodity prices were high as a result of wartime disruption and restocking needs after the war, and rose even further subsequently in 1950/51 as a result of the Korean war. Hence while the hypothesis is empirically supported if both the 1873–1938 period and the period since 1949–50 are considered separately, some doubts have been expressed about the postwar period and about the period since 1873 considered as a whole. However the doubts about the post-war period expressed by Spraos (1983) are only partial doubts; they only applied to the net barter terms of trade (NBTT) (and even there of doubtful validity) but vanished when looked at the 'Employment Corrected Double Factorial Terms of Trade' (ECDFTT). The double factorial terms of trade take the relative productivities into account as well as relative prices. The employment correction allows for the fact that the manufactured products from the North are produced under conditions of full employment, while the South was subject to chronic unemployment. (The first part of this correction would hardly be made for current measurement.) This shift to ECDFTT seems perfectly compatible with the Prebisch–Singer hypothesis since its main concern was with the welfare impact of terms of trade upon industrial and developing countries respectively which is a matter of productivity and employment as well as prices. Moreover, more detailed analysis of the post-war period or of the whole period since 1873 seems to confirm the hypothesis empirically even as far as simple NBTT are concerned. For example, Sapsford (1985) extended Spraos' analysis into the early eighties and applied statistical analysis to the whole series since 1900 to account for the wartime break and found that the Prebisch–Singer hypothesis was strongly borne out not only for the pre-war period since 1900 and the post-war period separately, but also for the whole period in spite of the wartime upward displacement. He determined the downward trend in the NBTT over the period 1900 to 1982 as 1.2% per annum. A.P. Thirlwall (1983, pp. 52–354) and Prabirjit Sarkar in a forthcoming paper on 'The terms of trade experience of Britain since the nineteenth century' also have no doubt about the genuineness and validity of the long-term declining trend in NBTT for primary commodity exports.

The empirical basis for a continuing post-war declining trend of terms of trade of developing countries or of primary exporters, in confirmation of the Prebisch–Singer hypothesis, can of course be taken as established only if oil prices are excluded. However, this exclusion of oil prices seems fully justified. The Prebisch–Singer hypothesis clearly refers to normal international market processes while the rise in oil prices was due to the application of producer power by a producer cartel in 1973 and again in 1979 to set aside market forces. In fact the need for such producer action and the need for international commodity agreements to raise and stabilize primary commodity prices is one of the possible policy conclusions arising from the Prebisch–Singer hypothesis. It could, of course, be taken as a weakness of Prebisch–Singer that it does not allow for such reaction to market pressures; but then the OPEC case has remained fairly isolated and it is by no means certain that market pressures will not in the end have the last word.

The underlying economic argument in explaining the trend towards deteriorating terms of trade observed and projected by Prebisch–Singer can be put under four headings:

(1) Differing elasticities of demand for primary commodities and manufactured goods. Primary commodities being inputs have a lower elasticity of demand because a 10 per cent drop (rise) in the price of the primary input will only mean a fractional drop in the price of the finished product – say 2 per cent instead of 10 per cent – and hence no great effect on demand can be expected. This means that in the case of a drop in prices there is no compensation in balance-of-payments terms (or 'income terms of trade') as a result of increasing volume. In the case of food the low price elasticity of demand is due to the fact that food is a basic need – and hence much of the income set free by a fall in the price of food will be devoted to other consumption goods rather than an increase in food consumption. Today the developing countries are net importers rather than exporters of food although this was not the case when the Prebisch–Singer hypothesis was developed in 1949/50. This low elasticity of demand, especially when combined with low elasticity of supply as emphasized by the classical analysis also means that there is great instability of primary commodity prices and hence terms of trade – both upward and downward. The Prebisch–Singer analysis did not always quite clearly distinguish the disadvantages of the present system of world trade for primary exporters due to price instability from those due to a deteriorating trend. In terms of instability, the Prebisch–Singer hypothesis was much more widely accepted and for example strongly anticipated by Keynes (1938), and also in the various memoranda and proposals by Keynes at the Bretton Woods conference aiming at an International Commodity Clearing House or even a world currency based on commodities.

(2) Demand for primary commodities is bound to expand less than demand for manufactured products. This is due partly to the lower income elasticity of demand for primary products, especially agricultural products (Engel's Law), and partly to the technological superiority of the industrial countries exporting manufactures. Part of that technological superiority is devoted to economies in the use of primary commodities and also to the development of synthetic substitutes for primary commodities. The latter has been a striking feature of economic development which has markedly accelerated since it was first emphasized by Prebisch–Singer (Singer 1950). The tendency towards balance of trade deficits for developing countries arising from such divergent demand trends will enforce currency depreciations which will introduce a further circle of terms of trade deterioration (although hopefully not of income terms of trade).

(3) The technological superiority of the industrial countries means that their exports embody a more sophisticated technology the control of which is concentrated in the exporting countries and especially in the large multinational firms located in those countries. This means that the prices of manufactured exports embody a Schumpeterian rent element for innovation and also a monopolistic profit element because of the size and power of multinational firms.

(4) The structure of both commodity markets and labour markets is different in industrial and developing countries. In the industrial 'centre' countries, labour is organized in trade unions and producers in strong monopolistic firms and producers' organizations, all very powerful at various times. This means that the results of technical progress and increased productivity are largely absorbed in higher factor incomes rather than lower prices for the consumers. In the developing 'peripheral' countries, to the contrary, where labour is unorganized, the rural surplus population (Lewis 1954) and its partial transfer into urban unemployment, open or disguised as explained in the Harris–Todaro model (Todaro 1969) make for a situation in which results of increased productivity are likely to

show in lower prices, benefiting the overseas consumer rather than the domestic producer. As long as we deal only with domestic production, such shifts in internal terms of trade between consumers and producers may not matter too much, partly because the two bodies are largely the same people, and partly because internal terms of trade can be influenced by domestic fiscal and other policies. But in international trade, the producers and consumers are in different countries; hence a tendency for productivity improvements mainly benefiting producers in the industrial countries but not in the developing countries will clearly affect terms of trade and international income distribution. Moreover, in many developing countries some of the major 'domestic' producers benefiting from higher productivity would be foreign investors and the higher profits flowing abroad would be equivalent in results to worsening terms of trade.

It will be noted that some of the four explanations for a deteriorating trend in terms of trade of developing countries relate as much or more to the characteristics of different types of *countries* – their different level of technological capacity, different organization of labour markets, presence or absence of surplus labour, etc. – as to the characteristics of different *commodities*. This indicates a general shift in the terms of trade discussion away from primary commodities *versus* manufactures and more towards exports of developing countries – whether primary commodities or simpler manufactures – *versus* the exports products of industrial countries – largely sophisticated manufactures and capital goods as well as skill-intensive services including technological know-how itself. As already mentioned, the initial hypothesis was formulated at a time when there was relatively little export of manufactures from developing countries. Since then there has been a considerable shift towards manufactures, including intensifying the export of primary commodities embodied in more highly processed manufactures. Although the early exponents of the Prebisch–Singer approach were often criticized for recommending import substituting industrialization (ISI) as a main policy conclusion, another equally logical policy conclusion would be export substituting industrialization (ESI) to get exports away from the deteriorating primary commodities. In fact this policy advice was given by some early followers of Prebisch–Singer to countries like India, where the possibilities of ESI seemed to exist at the time. However, the fact that some of the explanation for deteriorating terms of trade now relates to the characteristics of countries rather than commodities means that even ESI, a shift away from primary commodities to manufactures in the exports of developing countries has not disposed of the problem. The type of manufactures exported by developing countries in relation to the different types of manufactures exported by the industrial countries shared some of the disadvantages pointed out by Prebisch–Singer for primary commodities in relation to manufactures.

This can be demonstrated from some recent data. Taking trend equations for the period 1954–72 we find that in constant export unit values the prices of the primary commodities of developed countries fell by an annual average of 0.73%, but those of primary commodities of developing countries fell by 1.82% p.a. (both co-efficients significant at 1% level). This difference shows the existence of both commodity and country influences reinforcing each other. Similarly it can be shown that while the terms of trade for manufactures improved, they did so less for the manufactures of developing countries than those of industrial countries. Hence the deterioration in terms of trade of developing countries during this period can be attributed to three distinct factors:

(1) The rate of deterioration I prices of their primary commodities compared with those of primary commodities exported by industrial countries;
(2) a fall in prices of the manufactures exported by developing countries relative to the manufactures exported by industrial countries; and
(3) The higher proportion of primary commodities in the exports of developing countries which

means that the deterioration of primary commodities in relation to manufacturers affected them more than the industrial countries.

A quantitative weighting of these three factors is difficult but a broad estimate seems to show that they are of more or less equal importance. The original Prebisch–Singer hypothesis based on characteristics of commodities emphasized only the third factor, while the more recent formulations in terms of characteristics of countries include also the first two factors. It also shows that ESI mitigates the problem but does not entirely dispose of it. The shift in emphasis from commodity factors to country factors is particularly associated with the various theories of dependency (Prebisch and the work of the UN Economic Commission for Latin America (ECLA); Furtado 1964), of centre–periphery analysis (Seers 1983); and particularly of unequal exchange (Emmanuel 1972).

See Also

- ▶ Development Economics
- ▶ Periphery
- ▶ Structuralism

Bibliography

Ellsworth, P.T. 1956. The terms of trade between primary producing and industrial countries. *Inter-American Economic Affairs* 10: 47–65.

Emmanuel, A. 1972. *Unequal exchange*. New York/London: Monthly Review Press.

Furtado, C. 1964. *Development and underdevelopment*. Berkeley: University of California Press.

Haberler, G. 1961. Terms of trade and economic development. In *Economic development for Latin America*, ed. H.S. Ellis. London: Macmillan.

Keynes, J.M. 1938. The policy of government storage of foodstuffs and raw materials. *Economic Journal* 48: 449–460.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economics and Social Studies* 22: 139–191.

Morgan, T. 1959. The long-run terms of trade between agriculture and manufacturing. *Economic Development and Cultural Change* 8: 1–23.

Prebisch, R. 1950. *The economic development of Latin America and its principal problems*. New York: UN Economic Commission for Latin America.

Sapsford, D. 1985. The statistical debate on the net barter terms of trade between primary commodities and manufactures: A comment and some additional evidence. *Economic Journal* 95(379): 781–788.

Seers, D. 1983. *The political economy of nationalism*. Oxford: Oxford University Press.

Singer, H.W. 1950. The distribution of gains between investing and borrowing countries. *American Economic Review* 40: 473–485.

Spraos, J. 1980. The statistical debate on the net barter terms of trade between primary commodities and manufactures. *The Economic Journal* 90(357): 107–128.

Spraos, J. 1983. *Inequalising trade?* Oxford: Clarendon Press.

Spraos, J. 1985. A reply. *Economic Journal* 95(379): 789.

Thirlwall, A.P. 1983. *Growth and development, with special reference to developing economies*, 3rd ed. London: Macmillan.

Todaro, M.P. 1969. A model of labour migration and urban unemployment in less developed countries. *American Economic Review* 59(1): 138–148.

Viner, J. 1953. *International trade and economic development*. Oxford: Clarendon Press.

Terrorism, Economics Of

S. Brock Blomberg and Gregory D. Hess

Abstract

This article provides a comprehensive study of the economic determinants of domestic and transnational terrorism and the role that the economy plays in fostering a more peaceful world. We describe the research associated with the microfoundations of terrorist groups and how they organize. We also analyse models of conflict resolution to investigate the relative importance of macroeconomic factors for domestic and transnational terrorism. We describe a number of data-sets employed by researchers in the field and end by describing the most recent research which investigates the linkages between terrorism, democratization, globalization and development.

Keywords
Civil conflict; Democratization; Development; Foreign direct investment; Globalization; Gravity models; International trade; Terrorism, economics of; War and economics

JEL Classifications
E6; H1; H5; D74; O11

Since the prominent post-2000 terrorist incidents in high-income cities such as New York, Madrid and London, and the persistent incidence of terrorism in Middle East countries such as Israel and Iraq, both academia and the media have become involved in a careful examination of its causes. Terrorism is, however, neither new nor novel – indeed, the very origin of the term points to a long history, dating back to the late 1700s. (The word 'terrorism' apparently first appeared in the English language in reference to the 'reign of terror' associated with the rule of France by the Jacobins from 1793–94. The first incidence was actually reported in first century BC when Jewish terrorists, Zealots-Sicarri, incited a riot which led to a mass insurrection against the Roman Empire. See Laqueur 1977, pp. 7–8.) While terrorism has been present for longer than one might realize, research on the economics of terrorism has a shorter history. The literature has primarily focused on two areas: the micro-foundations of terrorism – understanding why organizations employ terrorist tactics – and the macroeconomic causes and consequences of these tactics. The latest wave of research highlights the relationship between terrorism, globalization and democratization.

Microeconomics of Terrorism

Microeconomic-based (or at least economic-sympathetic) research in terrorism generally concludes that terrorist organizations behave rationally (see, for example, Bueno de Mesquita 2005). So, while at key junctures terrorists initiate futile, high- cost (to terrorists' own selves) insurgencies and lose sight of whether the tactics really bring them closer to their political goals, for example, the behaviour is likely more driven by the clandestine nature of organizations and the daily struggle to maintain operational security by distorting rebels' understanding of the world outside their organization (see Bell 1990, 2002). This also points to the utility of propagating ideological movements in terrorism.

To whit, Hudson's review (1999) shows that for every study finding a purported psychological regularity, or social psychology, there exist a contradictory study. This is consistent with Krueger and Maleckova's (2003) finding that terrorists do not come from the poorer or uneducated segments of society. Given that organizations prefer to send competent terrorists, and thus select for the highest qualified, these results are best interpreted as showing that irrationality and individual oppression are not the underlying causes of terrorist incidents.

Moreover, if ideology plays a role in terrorist recruitment, its focus has clearly evolved. With respect to the spreading of revolutionary ideologies, there has been a change in the motive of terrorists since the November 1979 takeover of the US embassy in Tehran, the Iranian capital. Up to that point, terrorism had been primarily motivated by revolutionary and separatist ideologies – for example, the Red Army or Shining Path (see Wilkinson (2001). Since then, religious-based fundamentalism has played a more primary role. For example, the share of religious-based terrorist organizations had grown from four per cent to over 50 per cent by 1995 (Hoffman 1997).

However, it is misguided to search for the causes of terrorism in ideologies such as radical Islam. Rather, researchers such as Mishal and Sela (2002), Wilhelmsen (2004) and Brynjar and Hegghammer (2004) and Pape (2005) document that rational calculations rather than fundamentalist fury guide terrorist organizations' decision-making. From the perspective of leadership, the evidence is strong that terrorists strategically choose targets to best realize their political agenda. (In turn, Enders and Sandler 2002,

analyse the government's response to the innovations in the supply of terrorism.)

Macroeconomics of Terrorism

Given that terrorist behaviour can be described within a rational choice framework, many researchers have constructed models in order to evaluate the macroeconomic costs and consequences of terrorism based on Grossman (1991). Grossman provides the seminal economics paper investigating the integral linkages between civil conflict and the economy. He presents a general equilibrium model that treats insurrection and the suppression of insurrection as economic activities willingly undertaken by the participants. The ruler has to trade off higher taxes not only against the lower tax revenue that comes about when people devote less time to productive activities but also against the added cost of having to hire soldiering services to suppress insurrection. Grossman finds that economies in which the soldiering technology is effective can move themselves to no-conflict equilibria by devoting some resources to soldiering and keeping tax rates low.

With respect to the link between terrorism and the economic environment, Blomberg et al. (2004b) present a model that describes how one factor – the state of the economy – can lead groups to resort to terrorist attacks. Other authors such as Bernholz (2004) and Wintrobe (2002) have studied the important influences of increased fundamentalism and group solidarity in driving terrorist activity. Still, it is important to note that economic conditions help identify the underlying determinants of these activities. (There is also an existing literature that analyses how economics influences conflict in general. However, most of the analysis to this point has considered the impact on conflicts such as war without considering alternative types of conflict such as terrorism. For example, Hess and Orphanides 1995, 2001a, b, estimate the probability of conflict for the United States doubles when the economy has recently been in a recession and the president is running for re-election. Abadie and Gardeazabal 2003, also find a strong relation between the economy and terrorism.) The evidence is best summarized in Blomberg et al. (2004c) who provide an analysis of the relationship between economic growth phases (for example, expansions and contractions) and transitions into and out of terrorism incidents.

Violence, however, knows no bounds or lack of imagination. As such, it is often challenging to type forms of conflict into discrete, unbending categories. Acknowledging this point, there is a literature analysing the economic impact of terrorism as opposed to other forms of violence. Blomberg et al. (2004a), investigate the impact of various forms of conflict such as terrorism, internal wars and external wars on a country's economic growth. They find that, on average, the incidence of transnational terrorism has a significantly negative effect on growth, albeit one that is considerably smaller and less persistent than that associated with either external wars or internal conflict. They also find that terrorism is associated with a redirection of economic activity away from investment spending and towards government spending.

Other authors have concentrated on analysing one country in particular or on isolating one economic channel through which terrorism harms growth. Eckstein and Tsiddon (2004) provide an analysis of the macroeconomic consequences of terrorism in Israel, and find a large impact of domestic terrorism on economic activity. Using bilateral trade data, Blomberg and Hess (2006) establish that terrorism has a diminishing effect on international trade, and Blomberg and Mody (2006) demonstrate that violence also has a negative impact on foreign direct investment. Glick and Taylor (2004) also provide an analysis of the effect of external conflict on international trade over a longer historical period.

Terrorism Data and Empirical Regularities

To test theories of terrorism as well as to estimate the costs of terrorism, researchers need reliable measures of terrorism. Several competing international data-sets for terrorist incidents have been

cataloging attacks since the late 1960s. The dynamics across the major data-sets are roughly similar. In each data-set, the number of events increases during the period 1969 to 1987 (see the discussion in Blomberg and Hess 2007a, b). The US State Department and ITERATE data-sets estimate a similar steady increase from approximately 100 to 200 incidents per year up to 500 to 600 incidents per year. The RAND data-set estimates a similar trend, though the levels are smaller (from a base of approximately 100 incidents per year). The difference in the levels of terrorism in these data-sets arises because RAND does not include terrorism from state actor to non-state actor within a country and so systematically underestimates the number of attacks as compared with the other data-sets.

These trends demonstrate that the number of terrorist incidents steadily increased and peaked in the mid- to late 1980s. For several years thereafter, the worldwide intensity of transnational violence – violence motivated by international political considerations – fell steadily. In the late 1980s, according to the ITERATE data-set (see Mickolus et al. 2004), approximately one-and-one-half transnational violent events occurred every day. This frequency declined to less than one-half of an event a day by 2000. The decline also indirectly implies that the number of countries affected by a violent event fell over that period. This point has been more seriously addressed in Enders and Sandler (2005), who demonstrate that there has been no increase in violence from terrorism since the Al-Qaeda attacks on the United States on 11 September 2001. They show, if anything, that terrorism has fallen. However, over the time period in question the number of violent episodes may have risen – particularly since 11 September 2001. The average number of deaths per incident was 0.83 during 1968 to 1993. In seven of the following ten years, the number of deaths per incident was higher. Over the entire sample (1968–2003) there has been about one death per incident, and since 2001 the average has been five times that rate.

Blomberg and Hess (2006, 2007a, b) demonstrate two additional points. First, the recent drop in terrorism is systematic across regions, governments, income classes, and degrees of openness. Second, the hot spots for terrorism, as measured by incidence per capita, appear to be richer democracies, economies more open to trade and Middle Eastern countries.

Even though there is less systematic data available, terrorism was prominent before the 1960s and has evolved since its inception. In the late 19th and early 20th centuries, terrorists targeted political figures, the most notable examples being the 1881 assassination of Alexander II in Russia and the 1934 assassination of Alexander I of Yugoslavia. Current weapons and computer technology have transformed terrorists into literal artillery weapons causing a large amount of damage with only a few conspirators. Finally, many of these targets are now civilians rather than actual political heads.

Terrorism, Globalization, Democratization and Development

The changing dynamics of terrorism has led to an emerging set of papers investigating the relationship between globalization, democratization, development and terrorism. At its best, democracy provides a framework that aids the peaceful resolution of political conflicts (see Hess and Orphanides 2001a, for a discussion of how democracy affects the likelihood of external conflict). It offers access to decision makers and political institutions for citizens, while it also provides checks and balances within competing branches of the government. It also makes political organization cheaper and lowers the costs of political action. In turn, democracies should make illegal activities more expensive than legitimate political activity. In expectation, therefore, there should be less terrorist violence in democracies.

Alternatively, the key to the success of any terrorist act is recruitment and organization – both of which are made easier in free societies. Ironically, characteristics of democracies such as civil liberties and freedom of religion, association and movement could actually facilitate terrorist organization building. Moreover, as terrorists often target innocent civilians, free speech and a

free press (hallmarks of democracy) may be good channels for spreading fear. As a consequence, democracies could actually foster terrorist activities.

Is terrorism associated with the positive attributes of a democracy? Eubank and Weinberg (1994, 2001) find that terrorist groups are in fact more frequently hosted by democratic societies. Similarly, Li and Schaub (2004) find more incidents in democratic countries. Still, it may not be democracy that is the true driving force at work here; rather, the process of democratization may be the real culprit. For example, the experience of less democratic or newly democratizing countries such as Afghanistan and Iraq suggests that the transitional period between authoritarianism and democracy is a particularly susceptible one for terrorist activity (see Eubank and Weinberg 1998). Thus, the linkages between terrorism incidence and the evolution of a country's institutional governance are an important area for further study.

Other evidence on the link between democracy and transnational terrorism is decidedly mixed. Li (2005) attempts to disaggregate the many dimensions of democracy; he finds that voter turnout reduces terrorist incidents, but that constraints on government authority increase incidents. Finally, he finds press freedom raises incidents in a country. Taken as a whole, therefore, the effect of democracy on terrorism is not straightforward.

Globalization also affects the economic environment in which terrorist organizations can operate. If terrorism emerges from an environment of economic deprivation, then globalization, in so far as it enhances economic growth, may offset terrorist tendencies. Alternatively, if globalization increases inequality across countries and groups, then we might expect globalization to lead to more violence. Furthermore, globalization's associated lowered barriers to flows of goods, factors of production (including labour) and finance could make a network of terrorist operations cheaper to operate. Overall, globalization, like democracy, affects the costs, benefits and resources constraints of terrorists in many ways. Learning whether or not globalization is a net contributor to terrorism is therefore an empirical matter. Still, there is ample theory to support either conjecture.

Krug and Reinmoeller (2004) argue that globalization is an important determinant of terrorism. They build a model to explain the internationalization of terrorism as a natural response to a globalizing economy. As countries become more economically integrated and market-oriented, there is no discrimination between what certain terrorist groups might see as bad products and good products or investments. Moreover, the same advances in technology that allow for easy access of goods and services also allow for easy access to military hardware and technology. In the short run, globalization may have the consequence of creating a series of winners and losers. These same losers will find it easier to retaliate in response to their losses, thereby multiplying the effect of globalization on terrorism. (In a pooled cross-section analysis of globalization and transnational terrorism, Li and Schaub 2004, find that international trade and investment have little effect on the number of terrorist events.)

An alternative view put forth by Crenshaw (2001) is that it is naive to believe that globalization encourages international terrorism; while globalization and terrorism may seem to affect one another, there is something more complicated at work. She argues that the latest wave of terrorism should be seen as a series of civil wars which may be a strategically unified reaction to American power rather than directly to globalization.

In a series of papers, Blomberg and Hess (2006, 2007a) and Blomberg and Rosendorff (2007) are able to uncover several important linkages between globalization, democratization, development and terrorism by employing the workhorse model in the international trade literature – the gravity model. Terrorism is defined by the fact that 'its ramifications transcend national boundaries' (see Mickolus et al. 2004). Transnational terrorism requires, therefore, a flow of resources across borders, and these papers consider both the source and target countries' characteristics in determining terrorism. The characteristics of a country that might make it a likely target country may indeed be very different from the characteristics that make a country a likely source of international terrorism. The features of the polity that make a country a terrorist-producer may be different from the political

structures, institutions and environment that make a state a terrorist target. The conclusion from these papers is that the advent of democratic institutions, high income and more openness in a *source* country significantly reduces conflict. However, the advent of these same positive developments in *targeted* countries actually increases conflict. *Ceteris paribus*, the impact of being a democracy or participating in the WTO/IMF for a *source* country decreases the number of terrorist strikes by about two or three per year, which is more than two standard deviations greater than the average number of strikes between any two countries in a given year.

Terrorism is a nascent field in economics. Sadly, though realistically, it is also likely to be a growing field in economics. Conflict and the economy are intimately related at the domestic and at the international levels, and across various forms of governance. Moreover, innovations in the strategy of terrorism will continue (see, for example, Berman and Laitin's analysis, 2005, of suicide attacks). Terrorism therefore will remain part of the violence spectrum and affect our economic landscape for years to come.

Bibliography

Abadie, A., and J. Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque country. *American Economic Review* 93: 113–132.

Bell, J.B. 1990. Revolutionary dynamics: The inherent inefficiencies of the underground. *Terrorism and Political Violence* 2: 193–211.

Bell, J.B. 2002. The organization of Islamic terror: The global jihad. *Journal of Management Inquiry* 11: 261–266.

Berman, E., and D. Laitin. 2005. *Hard targets: Theory and evidence on suicide attacks*, Working Paper No. 11740. Cambridge, MA: NBER.

Bernholz, P. 2004. Supreme values as the basis for terror. *European Journal of Political Economy* 20: 317–333.

Blomberg, S.B., and G.D. Hess. 2006. How much does violence tax trade? *Review of Economics and Statistics* 88: 599–612.

Blomberg, S.B., and G.D. Hess. 2007a. The Lexus and the olive branch. In *Economic consequences of terrorism in developed and developing countries*, ed. P. Keefer and N. Loayza. Cambridge: Cambridge University Press.

Blomberg, S.B., and G.D. Hess. 2007b. From (no) butter to guns? Understanding the economic role in terrorism. In *Economic consequences of terrorism in developed and developing countries*, ed. P. Keefer and N. Loayza. Cambridge: Cambridge University Press.

Blomberg, S.B., G.D. Hess, and A. Orphanides. 2004a. The macroeconomic consequences of terrorism. *Journal of Monetary Economics* 51: 1004–1030.

Blomberg, S.B., G.D. Hess, and A. Weerapana. 2004b. An economic model of terrorism. *Conflict Management and Peace Science* 21: 17–28.

Blomberg, S.B., G.D. Hess, and A. Weerapana. 2004c. Economic conditions and terrorism. *European Journal of Political Economy* 20: 463–478.

Blomberg, S.B., and A. Mody. 2006. *How severely does violence deter international investment?* Working paper, International Monetary Fund.

Blomberg, S.B., and P. Rosendorff. 2007. A gravity model of globalization, democracy and transnational terrorism. In *Guns and butter: The economic causes and consequences of violent conflict*, ed. G. Hess. Cambridge, MA: MIT Press.

Brynjar, L., and T. Hegghammer. 2004. Jihadi strategic studies: The alleged Al Qaida policy study preceding the Madrid bombings. *Studies in Conflict and Terrorism* 27: 355–375.

Bueno de Mesquita, E. 2005. The quality of terror. *American Journal of Political Science* 49: 515–530.

Crenshaw, M. 2001. Why America? The globalization of civil war. *Current History* 100: 425–432.

Eckstein, Z., and D. Tsiddon. 2004. Macroeconomic consequences of terror: Theory and the case of Israel. *Journal of Monetary Economics* 51: 971–1002.

Enders, W., and T. Sandler. 2002. Patterns of transnational terrorism, 1970–1999: Alternative time-series estimates. *International Studies Quarterly* 46: 145–165.

Enders, W., and T. Sandler. 2005. After 9–11: Is it all so different now? *Journal of Conflict Resolution* 49: 259–277.

Eubank, W., and L. Weinberg. 1994. Does democracy encourage terrorism? *Terrorism and Political Violence* 6: 417–443.

Eubank, W., and L. Weinberg. 1998. Terrorism and democracy: What recent events disclose. *Terrorism and Political Violence* 10: 108–118.

Eubank, W., and L. Weinberg. 2001. Terrorism and democracy: Perpetrators and victims. *Terrorism and Political Violence* 13: 155–164.

Glick, R., and A. Taylor. 2004. *Collateral damage: The economic impact of war*, Working Paper No. 11565. Cambridge, MA: NBER.

Grossman, H.I. 1991. A general equilibrium model of insurrections. *American Economic Review* 81: 912–921.

Hess, G.D., and A. Orphanides. 1995. War politics: An economic, rational-voter framework. *American Economic Review* 85: 828–846.

Hess, G.D., and A. Orphanides. 2001a. Economic conditions, elections, and the magnitude of foreign conflicts. *Journal of Public Economics* 80: 121–140.

Hess, G.D., and A. Orphanides. 2001b. War and democracy. *Journal of Political Economy* 109: 776–810.
Hoffman, B. 1997. The confluence of international and domestic trends in terrorism. *Terrorism and Political Violence* 9: 1–15.
Hudson, R.A. 1999. *The sociology and psychology of terrorism: Who becomes a terrorist and why?* Washington, DC: Federal Research Division, Library of Congress.
Krueger, A.B., and J. Maleckova. 2003. Education, poverty and terrorism: Is there a causal connection? *Journal of Economic Perspectives* 17(4): 119–144.
Krug, B., and P. Reinmoeller. 2004. *The hidden cost of ubiquity: Globalisation and terrorism*. Research paper, Erasmus Research Institute of Management, RSM Erasmus University.
Laqueur, W. 1977. *Terrorism*. London: Weidenfeld and Nicolson.
Li, Q. 2005. Does democracy promote or reduce transnational terrorist incidents? *Journal of Conflict Resolution* 49: 278–297.
Li, Q., and D. Schaub. 2004. Economic globalization and transnational terrorism: A pooled time-series analysis. *Journal of Conflict Resolution* 48: 230–258.
Mickolus, E.F., T. Sandler, J.M. Murdock, and P. Flemming. 2004. *International terrorism: Attributes of terrorist events, 1968–2003 (ITERATE 5)*. Dunn Loring: Vinyard Software.
Mishal, S., and A. Sela. 2002. Participation without presence: Hamas, the Palestinian authority and the politics of negotiated existence. *Middle Eastern Studies* 38(3): 1–26.
Pape, R. 2005. *Dying to win*. New York: Random House.
Wilhelmsen, J. 2004. *When separatists become Islamists: The case of Chechnya*, Report No. 2004/00445. Kjeller: Norwegian Defence Research Establishment (FFI).
Wilkinson, P. 2001. *Terrorism versus democracy: The liberal state response*. London: Frank Cass.
Wintrobe, R. 2002. Can suicide bombers be rational? Unpublished manuscript, University of Western Ontario.

Testing

Frank Kleibergen

Abstract

Hypothesis testing is the customary instrument for analysing the empirical validity of an economic theory. Hypothesis testing is thus an important tool for conducting statistical inference in economic models. In this article we show how an economic theory is tested in a statistical model. We begin with the discussion of the basic results on hypothesis testing and then focus on some recent developments that have improved testing in commonly used economic models such as the linear instrumental variables regression model. We use a real economic example to illustrate the main findings.

Keywords

Anderson–Rubin statistic; Bootstrap; Generalized method of moments; Hypothesis testing: *see* testing; Lagrange multipliers; Least squares; Likelihood ratios; Limited information maximum likelihood; Linear models; Maximum likelihood; Neymann–Pearson Lemma; Price elasticity; Probability; Statistical inference; Testing; Two-stage least squares; Wald statistics

JEL Classifications

C12

Hypothesis testing is the customary instrument for analysing the empirical validity of an economic theory. This theory is reduced to a hypothesis which is tested in a statistical model. Hypothesis testing is thus an important tool for conducting statistical inference in economic models. An impressive literature has emerged which discusses tests of economic hypotheses. Instead of providing an incomplete overview of this literature, we provide a somewhat hands on discussion of testing in which we show how an economic theory is tested in a statistical model. We therefore begin with the discussion of the basic results on hypothesis testing and later focus on some recent developments that have improved testing in commonly used economic models. We use a real economic example to illustrate the main findings.

When testing an economic hypothesis, we want our test results to hold generally and not to be affected by highly specific assumptions on the statistical model such as, for example, the distribution of the disturbances. Under these general conditions, the finite sample distributions of the

involved test statistics are unknown. The realized values of the test statistics are then confronted with critical values that result from the large sample distributions of the statistics under the hypothesis of interest.

Since this is currently the common approach to testing, our discussion is conducted solely from the large sample perspective.

We illustrate the tests of an economic theory by testing for a unit demand price elasticity for the demand for oranges. We use data on the demand for oranges in the United States during 1910–59. The data results from Nerlove and Waugh (1961) and is also used in Berndt (1991, pp. 417–20). The demand equation is specified as

$$\log(P_t) = \alpha + \gamma \log(Q_t) + \beta \log(RI_t) + \mathcal{E}_t, \quad t = 1, \ldots, T, \tag{1}$$

with Q_t the traded quantity, P_t the price of oranges and RI_t the real disposable income. The other available series are current (AC_t) and past advertisement (AP_t, averaged over the last ten years). We test the hypothesis of a unit demand price elasticity $H_0 : \gamma = -1$ against the alternative hypothesis $H_1 : \gamma \neq -1$.

The Trinity of Tests

Most tests result from one of the three main principles for constructing a test statistic– Wald, likelihood ratio (LR) and Lagrange multiplier (LM) or score – which are often referred to as the trinity of tests (for example, Engle 1984). When $p(y{:}\theta)$ is the joint density of the $T \times 1$ data vector y and we want to test the hypothesis $H_0 : \theta = \theta_0$ on the $m \times 1$ vector of parameters θ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, these three statistics read:

$$\begin{aligned} \text{Wald}(\theta_0) &= T(\hat{\theta} - \theta_0)' V\left(\hat{\theta}\right)^{-1} \left(\hat{\theta} - \theta_0\right) \\ \text{LR}(\theta_0) &= 2[L\left(y; \hat{\theta}\right) - L(y; \theta_0)] \\ \text{LM}(\theta_0) &= \frac{1}{T} s(y; \theta_0)' I(\theta_0)^{-1} s(y; \theta_0), \end{aligned} \tag{2}$$

with $L(y; \theta)$ the logarithm of the likelihood or joint density $p(y; \theta)$, $L(y; \theta) = \log(p(y; \theta))$; $s(y; \theta)$ is the score, $s(y; \theta) = \frac{\partial}{\partial \theta} L(y; \theta)$; $\hat{\theta}$ is the maximum likelihood estimator under H_1 so $s\left(y; \hat{\theta}\right) = 0$, $V(\theta)$ is the covariance matrix of $\hat{\theta}$ and $I(\theta)$is the information matrix ($V(\theta) = I(\theta)^{-1}$),

$$I(\theta) = -E\left[\frac{1}{T} \frac{\partial^2}{T \partial\theta \partial\theta'} L(y; \theta)\right] \tag{3}$$

The three statistics provide different measures of the relative distance between H_0 and H_1. Under sufficient regularity conditions which ensure the consistency and asymptotic normality of $\hat{\theta}$, all three statistics converge under H_0 to the same $\chi^2(m)$ distributed random variable when the sample size becomes large (for example, Newey and MacFadden 1994, Th. 9.2). When these regularity conditions hold, usage of a specific statistic is typically a matter of computational ease. The Wald and LM statistics analyse the model only under H_1 or H_0 resp. so one could have a preference for either, given the computational effort to analyse the model under H_0 or H_1. The LR statistic involves the analysis of the model under H_0 and H_1 and is therefore more demanding to compute than the Wald and LM statistics. When one is conducting tests on only one element of θ, the so-called *t*-statistic is often used which equals the square root of the Wald statistic and has a large sample normal distribution under H_0.

Significance, Size and Power

When we test H_0, we specify a significance level of $100 \times (1 - \alpha)\%$ which sets the probability that we reject H_0 while it is true equal to α. The critical value associated with this significance level is then such that the probability mass of the large sample distribution of the statistic under H_0 above the critical value equals α. We then reject H_0 with $(1 - \alpha) \times 100\%$ significance when the realized value of the statistic exceeds the $(1 - \alpha) \times 100\%$ critical value.

Another manner to test H_0 with $100 \times (1 - \alpha)\%$ significance is by using the

p-value associated with the realized value of the statistic. The p-value equals the probability mass of the large sample distribution of the statistic under H_0 that lies above the realized value of the statistic. Hence, we reject H_0 with $100 \times (1 - \alpha)\%$ significance when the p-value is less than α.

Tests of $H_0 : \theta = \theta_0$ with $100 \times (1 - \alpha)\%$ significance for a range of values of θ_0 can be used to obtain the $100 \times (1 - \alpha)\%$ confidence set of θ. The $100 \times (1 - \alpha)\%$ confidence set of θ contains all values of θ_0 for which H_0: $\theta = \theta_0$ is not rejected with $100 \times (1 - \alpha)\%$ significance and therefore contains the true value of θ with probability $1-\alpha$.

Besides computational issues there are several other reasons to prefer a specific statistic, especially when it is unclear whether the regularity conditions, which imply the large sample distribution of the statistic under H_0, hold. Examples of such other reasons are invariance to transformations of the parameters, observed size of the statistics and discriminatory power.

Especially in models that are nonlinear in the parameters, it is appealing to use a statistic whose specification is invariant to nonlinear transformations of the parameters so it does not depend on the specification of the model. This property is violated by the Wald statistic but satisfied by the LR and LM statistics (for example, Dagenais and Dufour 1991; Dufour 1997). Hence it is better to use either the LR or LM statistic in such models.

The specification of the significance level of the test intends to control the *Type I error* or probability that we reject H_0 while it holds. The rejection frequency under H_0, to which we refer as the size of the test, should therefore coincide with α.

Because we use the large sample distribution of the statistic under H_0 instead of the unknown finite sample distribution to obtain the critical value of the test, this is, however, not the case. The statistic whose size properties dominate those of the others is then typically preferred. The size properties of the different statistics can often be improved by computing the critical values using the bootstrap instead of the large sample distribution (for example, Horowitz 2001).

The *Type II error* of the statistic is the probability of not rejecting H_0 while it is false. We thus prefer statistics that minimize the *Type II error*, or, put differently, maximize the discriminatory power while preserving an adequate size. When the likelihood function is known, the Neymann–Pearson Lemma implies that the LR statistic is the most powerful statistic for testing a point null hypothesis, like $H_0 : \theta = \theta_0$, against a point alternative, like $H_2 : \theta = \theta_2$. For composite alternatives, like $H_1 : \theta \neq \theta_0$, there is typically no statistic that is the most powerful one in all cases.

Tests in the Linear Regression Model

To construct test statistics for the linear demand Eq. (1), we assume that the disturbances are independently and identically distributed so we can estimate the parameters using least squares:

$$\log(P_t) = -6.19_{(1.66)} - 0.79_{(0.11)}\log(Q_t) + 0.92_{(0.23)}\log(RI_t) + \hat{\mathcal{E}}_t \tag{4}$$

(The standard errors are reported below the parameter estimates.) To compute the Wald, LR and LM statistics that test for a unit demand price elasticity, we further assume the disturbances to be independently identically normally distributed with mean zero. The specifications of the three tests then read

$$\begin{aligned} \text{Wald}(\theta_0) &= \frac{\text{RSSR} - \text{USSR}}{\text{USSR}/T}, \\ \text{LR}(\theta_0) &= T\log\left[\frac{\text{RSSR}}{\text{USSR}}\right] \text{and} \\ \text{LM}(\theta_0) &= \frac{\text{RSSR} - \text{USSR}}{\text{RSSR}/T}, \end{aligned} \tag{5}$$

which test $H_0 : \theta = \theta_0$ and where USSR is the unrestricted sum of squared residuals $\hat{\mathcal{E}}_t$, that is, the residuals under H_1, and RSSR is the restricted sum of squared residuals, that is, the residuals under H_0. Under H_0 and when the disturbances are independently identically distributed with mean zero and a finite fourth order moment

all three statistics converge to the same $\chi^2(1)$ distributed random variable when the sample size gets large.

The expression of the LM statistic in (5) is such that we can compute it as well using an auxiliary regression of the restricted residuals under H_0 on all explanatory variables. The expression of $LM(\theta_0)$ is then such that it equals T times the R^2 of this regression.

The values of the statistics that test for a unit demand price elasticity that result from (5) read

$$\begin{aligned} \text{Wald}(-1) = 3.62 &> \text{LR}(-1) = 3.50 \\ &> \text{LM}(-1) = 3.38 \end{aligned} \tag{6}$$

(The value of the Wald statistic in (6) is computed using (5) but could alternatively have been computed using the least squares estimator and standard error that are reported in (4) since $3.62 \approx \left(\frac{-0.79+1}{0.11}\right)^2$). All three statistics are smaller than the 95 per cent critical value of 3.84 that results from the large sample distribution of the statistics under H_0, which is the $\chi^2(1)$ distribution, so we do not reject the unit demand price elasticity with 95 per cent significance. The Wald statistic that tests for a unit demand price elasticity exceeds the value of the LR statistic which again exceeds the value of the LM statistic. This result always holds for tests on the parameters of the linear regression model and is not a result of the involved data.

Specification Tests

Estimation of the demand Eq. (1) by least squares as in (4) presumes that the traded quantity is exogenous since least squares leads to an inconsistent estimator when the traded quantity is endogenous. When the traded quantity is endogenous, we need to use an estimator that remains consistent in that case, like, for example, the two-stage least squares (2SLS) estimator or the limited information maximum likelihood (LIML) estimator (see, for example, Theil 1953; Hood and Koopmans 1953).

Exogeneity or endogeneity of the traded quantity lead to different specifications of the statistical model for the demand for oranges. A test for the appropriate specification of the model is the Durbin–Wu–Hausman (DWH) statistic, which tests the difference between two estimators, $\tilde{\theta}$ and $\hat{\theta}$, one of which, $\hat{\theta}$, is efficient and consistent in the model under the null hypothesis but not in the model under the alternative hypothesis, while the other estimator, $\tilde{\theta}$, is consistent in both models (see Durbin 1954; Wu 1973; Hausman 1978):

$$DWH(\theta) = T\left(\tilde{\theta} - \hat{\theta}\right)' \left[V\left(\tilde{\theta}\right) - V\left(\hat{\theta}\right)\right]^{-} \left(\tilde{\theta} - \hat{\theta}\right) \tag{7}$$

with $V\left(\tilde{\theta}\right)$ and $V\left(\hat{\theta}\right)$ the covariance matrices of $\tilde{\theta}$ and $\hat{\theta}$ and $[\ldots]^{-}$ the generalized inverse operator. Under sufficient regularity conditions, the DWH statistic converges under H_0 to a $\chi^2(m)$ distributed random variable with m the minimum of the number of elements of θ and the rank of the matrix $V\left(\tilde{\theta}\right) - V\left(\hat{\theta}\right)$.

Using the current and past advertisement variables as instruments, we computed the DWH statistic to test the null hypothesis of exogeneity of the traded quantity against the alternative hypothesis of endogeneity using both the 2SLS and LIML estimators:

$$\begin{aligned} \text{DWH}_{\text{2SLS}}(\theta) &= T\left(\tilde{\theta}_{\text{2SLS}} - \hat{\theta}_{\text{LS}}\right)' \left[V\left(\tilde{\theta}_{\text{2SLS}}\right) - V\left(\hat{\theta}_{\text{LS}}\right)\right]^{-} \left(\tilde{\theta}_{\text{2SLS}} - \hat{\theta}_{\text{LS}}\right) = 4.99, \\ \text{DWH}_{\text{LIML}}(\theta) &= T\left(\tilde{\theta}_{\text{LIML}} - \hat{\theta}_{\text{LS}}\right)' \left[V\left(\tilde{\theta}_{\text{LIML}}\right) - V\left(\hat{\theta}_{\text{LS}}\right)\right]^{-} \left(\tilde{\theta}_{\text{LIML}} - \hat{\theta}_{\text{LS}}\right) = 4.96. \end{aligned} \tag{8}$$

Both statistics exceed the 95 per cent critical value of 3.84 of the large sample distribution of the DWH statistic under H_0, which is a χ^2-(1) distribution. Hence, we reject with 95 per cent significance that the traded quantity is exogenous. This implies that we have to account for the endogeneity of the traded quantity when we test the unit demand price elasticity hypothesis.

Tests in the Linear Instrumental Variables Regression Model

To accommodate the endogeneity of the traded quantity, we test the demand price elasticity in a linear instrumental variables regression model

$$y_t = x_t\beta + w_t'\gamma + \varepsilon_t \; x_t = z_t'\pi + w_t'\delta + v_t \qquad (9)$$

where y_t and x_t are the endogenous variables, w_t is a $k_w \times 1$ vector that contains the included exogenous variables, z_t is a $k_s \times 1$ vector that contains the instruments and ε_t and v_t are the disturbances. The instruments z_t are uncorrelated with the structural ε_t disturbances ε_t. We assume that the vectors of disturbances $\begin{pmatrix} \varepsilon_t \\ v_t \end{pmatrix}, t = 1, \ldots, T$ are independently identically distributed. The variables for the demand for oranges are such that $y_t = \log(P_t), \quad x_t = \log(Q_t), \; w_t = \begin{pmatrix} 1 \\ \log(RI_t) \end{pmatrix}$ and $z_t = \begin{pmatrix} \log(AC_t) \\ \log(AP_t) \end{pmatrix}$ structural parameter β is typically our parameter of interest and we therefore partial out w_t from the model by replacing all remaining variables with the residuals that result from regressing them on w_t:

$$\hat{y}_t = \hat{x}_t\beta + \varepsilon_t \; \hat{x}_t = \hat{z}_t'\pi + v_t \qquad (10)$$

with $\hat{y}_t, \hat{x}_t$ and $\hat{z}_t$ the residuals that result from regressing y_t , x_t and z_t resp. on w_t.

We want to test a hypothesis on the structural parameter β, $H_0 : \beta = \beta_0$, like, for example, that of a unit demand price elasticity. We discuss some statistics that can be used for this purpose most of which belong to the trinity of tests.

Wald Statistics

Using either the 2SLS or LIML estimator, we can test H_0 using a Wald statistic:

$$\begin{aligned} \text{Wald}_{\text{2SLS}}(\beta_0) = \quad & T\left(\hat{\beta}_{\text{2SLS}} - \beta_0\right)' V\left(\hat{\beta}_{\text{2SLS}}\right)^{-1} \\ & \left(\hat{\beta}_{\text{2SLS}} - \beta_0\right) \\ \text{Wald}_{\text{LIML}}(\beta_0) = \quad & T\left(\hat{\beta}_{\text{LIML}} - \beta_0\right)' V\left(\hat{\beta}_{\text{LIML}}\right)^{-1} \\ & \left(\hat{\beta}_{\text{LIML}} - \beta_0\right), \end{aligned} \qquad (11)$$

with $\hat{\beta}_{\text{2SLS}}$ the 2SLS estimator:

$$\hat{\beta}_{2SLS} = (\hat{\pi}' \sum_{t=1}^{T} \hat{z}_t \hat{x}_t)^{-1} \hat{\pi}' \sum_{t=1}^{T} \hat{z}_t \hat{y}_t,$$
$$\hat{\pi} = (\sum_{t=1}^{T} \hat{z}_t \hat{z}_t')^{-1} \sum_{t=1}^{T} \hat{z}_t \hat{x}_t \text{ and } \hat{\beta}_{\text{LIML}} \text{ the LIML}$$

estimator:

$$\hat{\beta}_{\text{LIML}} = \text{argmin}_\beta \frac{\left[\sum_{t=1}^{T} \hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]' \left[\sum_{t=1}^{T} \hat{z}_t\hat{z}_t'\right]^{-1} \left[\sum_{t=1}^{T} \hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]}{\sum_{t=1}^{T} (\hat{y}_t - \hat{x}_t\beta)^2 - \left[\sum_{t=1}^{T} \hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]' \left[\sum_{t=1}^{T} \hat{z}_t\hat{z}_t'\right]^{-1} \left[\sum_{t=1}^{T} \hat{z}_t(\hat{y}_t - \hat{x}_t\beta\right]} \qquad (12)$$

Both Wald statistics converge under H_0 and a number of regularity conditions which rule out zero values of π to a $\chi^2(1)$ distributed random variable when the sample size gets large (for example, Newey and MacFadden 1994). The assumption of a non-zero value of π for the large

sample distribution implies that finite sample distributions of both Wald statistics depend on the value of π. The actual size of the Wald statistics can therefore deviate considerably from the assumed *Type 1 error* which makes these statistics unreliable for usage in practice (for example, Nelson and Startz 1990; Bound et al. 1995; Dufour 1997; Staiger and Stock 1997). The bootstrap cannot be used to overcome these size distortions (for example, Horowitz 2001).

Anderson–Rubin Statistic

Anderson and Rubin (1949) construct a test for H_0 by substituting the equation of $\hat{x}_t$ into the equation of $\hat{y}_t$:

$$\hat{y}_t - \hat{x}_t\beta_0 = \hat{z}_t'\phi + u_t \tag{13}$$

with $\phi = \pi(\beta - \beta_0)$ and $u_t = \varepsilon_t + v_t(\beta - \beta_0)$. Under H_0, ϕ equals zero and a test for H_0 can therefore be conducted using a test for a zero value of ϕ. Anderson and Rubin (1949) proposed to use the F-statistic that tests for a zero value of φ in (13) for this purpose. This F-statistic is commonly referred to as the Anderson–Rubin (AR) statistic.

When the disturbances are independently and identically distributed with finite fourth-order moments, the AR statistic converges under H_0 to a $\chi^2(k_z)/k_z$ distributed random variable when the sample size gets large. This large sample distribution of the AR statistic does not depend on the value of π, which makes the AR statistic a more reliable statistic for practical purposes than the Wald statistics in (11). A disadvantage of the AR statistic is that its large sample distribution is proportional to a χ^2 distribution with a degrees of freedom parameter that equals the number of instruments while we conduct a test on only one parameter. This reduces the discriminatory power of the AR statistic when the number of instruments is large, which is often the case.

LM Statistic

When we assume the vector of disturbances $\begin{pmatrix}\varepsilon_t \\ v_t\end{pmatrix}$ to be independently identically normally distributed with mean zero, we can construct the likelihood function and therefore also the LM statistic for testing H_0 (see Kleibergen 2002):

$$LM(\beta_0) = \frac{1}{\widetilde{\sigma}_{\varepsilon\varepsilon}}\left(\sum_{t=1}^{T}\hat{z}_t\widetilde{\varepsilon}_t\right)'\widetilde{\pi}(\beta_0) \left[\widetilde{\pi}(\beta_0)\prime\left(\sum_{t=1}^{T}\hat{z}_t\hat{z}_t'\right)\widetilde{\pi}(\beta_0)\right]^{-1}\left(\sum_{t=1}^{T}\hat{z}_t\widetilde{\varepsilon}_t\right) \tag{14}$$

with $\widetilde{\varepsilon}_t = \hat{y}_t - \hat{x}_t\beta_0$,

$$\widetilde{\pi}(\beta_0) = \left(\sum_{t=1}^{T}\hat{z}_t\hat{z}_t'\right)^{-1}\sum_{t=1}^{T}\hat{z}_t\left[\hat{x}_t - \widetilde{\varepsilon}_t\frac{\widetilde{\sigma}_{v\varepsilon}}{\widetilde{\sigma}_{\varepsilon\varepsilon}}\right],$$
$$\widetilde{\sigma}_{\varepsilon\varepsilon} = \frac{1}{T-k_z}\sum_{t=1}^{T}(\tilde{y}_t - \tilde{x}_t\beta_0)^2,$$
$$\widetilde{\sigma}_{v\varepsilon} = \frac{1}{T-k_z}\sum_{t=1}^{T}\tilde{x}_t(\tilde{y}_t - \tilde{x}_t\beta_0) \tag{15}$$

and where $\tilde{x}_t$ and $\tilde{y}_t$ are the residuals that result from regressing x_t and y_t on w_t and z_t. When the disturbances $\begin{pmatrix}\varepsilon_t \\ v_t\end{pmatrix}$ are independently identically distributed and have finite fourth order moments, the LM statistic converges to a $\chi^2(1)$ distributed random variable when the sample size increases. The convergence of the LM statistic does not depend on the value of π. The large sample distribution of the LM statistic under H_0 is therefore typically a rather accurate approximation of the finite sample distribution, and this approximation can even be further improved by using the bootstrap (see Kleibergen 2004).

LR Statistic

Under identically independently normal distributed disturbances, Moreira (2003) constructs the likelihood ratio statistic to test H_0:

$$\mathrm{LR}(\beta_0) = \frac{1}{2}\Big[\mathrm{AR}(\beta_0) - \mathrm{r}(\beta_0) + \sqrt{\mathrm{AR}(\beta_0) + \mathrm{r}(\beta_0))^2 - 4\mathrm{r}(\beta_0)(\mathrm{AR}(\beta_0) - \mathrm{LM}(\beta_0))}\Big] \tag{16}$$

T

with AR(β_0) k_w times the AR statistic that tests H_0:

$$AR(\beta_0) = \frac{1}{\widetilde{\sigma}_{\varepsilon\varepsilon}} \left(\sum_{t=1}^{T} \hat{z}_t \widetilde{\varepsilon}_t \right)' \left(\sum_{t=1}^{T} \hat{z}_t \hat{z}_t' \right)^{-1} \left(\sum_{t=1}^{T} \hat{z}_t \widetilde{\varepsilon}_t \right) \tag{17}$$

and r(β_0) a statistic that tests for a zero value of π under H_0, so by using $\widetilde{\pi}(\beta_0)$,

$$r(\beta_0) = \frac{1}{\widetilde{\sigma}_{vv.\varepsilon}} \widetilde{\pi}(\beta_0)' \left(\sum_{t=1}^{T} \hat{z}_t \hat{z}_t' \right) \widetilde{\pi}(\beta_0) \tag{18}$$

where $\widetilde{\sigma}_{vv.\varepsilon} = \widetilde{\sigma}_{vv} - \frac{\widetilde{\sigma}_{v\varepsilon}^2}{\widetilde{\sigma}_{.\varepsilon\varepsilon}}$ and $\widetilde{\sigma}_{vv} = \frac{1}{T-k_z} \sum_{t=1}^{T} \tilde{x}_t^2$. Moreira (2003) shows that, when the disturbances are independently identically distributed with finite fourth-order moments, the large sample distribution of LR(β_0) under H_0 is conditional on the value of r(β_0). We thus need to use a different critical value to determine the significance of a realized value of the LR statistic for every value of r(β_0). When r (β_0) is zero, the large sample distribution of LR(β_0) is identical to a $\chi^2(k_w)$ distribution while it equals the $\chi^2(1)$ zdistribution for large values of r(β_0). Besides the dependence on r(β_0), the large sample distribution of LR(β_0) does not depend on π, which makes LR(β_0) a trustworthy statistic for practical purposes. Andrews et al. (2006) show that the LR statistic is the most powerful of those statistics whose large sample distributions do not depend on π and are invariant under transformations of the model.

The Unit Demand Price Elasticity

We test for a unit demand price elasticity using each of the above statistics.

$$\begin{aligned} &\text{Wald}_{2SLS}(-1) = 191 \; \text{Wald}_{LIML}(-1) = 178 \\ &AR(-1) = 73.7 \; LM(-1) = 67.3 \; LR(-1) = 69.1. \end{aligned} \tag{19}$$

The value of r(β_0) is 174, which makes the large sample distribution of LR(−1) given r(β_0) identical to the large sample distribution of LM (−1), which is a $\chi^2(1)$. The large sample distribution of AR(−1) is a $\chi^2(2)$ distribution, and the large sample distributions of the Wald statistics are $\chi^2(1)$ distributions while a non-zero value of π is assumed.

All statistics reject the hypothesis of a unit demand price elasticity with 95 per cent significance. This shows the importance of accounting for the endogeneity of the traded quantity since this hypothesis was not rejected in the linear regression model that assumes the traded quantity to be exogenous. The values of the statistics whose large sample distributions are not influenced by the value of π, that is, AR(− 1), LM(− 1) and LR(− 1), are all of the same order of magnitude, while the Wald statistics are much larger. This indicates the different behaviour of these statistics and we recommend that these Wald statistics not be used.

More General Specifications

We discussed the trinity of tests in a linear model estimated using either least squares or instrumental variables. The Wald, LM and LR statistics extend to a large variety of models which are possibly nonlinear in the parameters and have unknown likelihood functions. The expression of the Wald statistic is such that it can be applied to any estimator which has a normal large sample distribution and for which a consistent estimator of the asymptotic variance exists. The LM test is applicable in any model where the estimators solve a first-order condition. The LR test is based on the difference of an objective function under H_0 and H_1, a specification that allows it to accommodate more general statistical models (for example, Engle 1984; Newey and MacFadden 1994). In these general settings, such as, for example, the generalized method of moments, the large sample distribution of the Wald statistic is often affected by nuisance parameters while the large sample distributions of the LM and LR statistics remain robust to the value of these nuisance parameters (for example, Kleibergen 2005). Hence, the LM and LR statistics often provide more reliable tests.

See Also

- Bootstrap
- Endogeneity and Exogeneity
- Functional Central Limit Theorems
- Instrumental Variables
- Simultaneous Equations Models
- Statistical Inference
- Statistics and Economics

Bibliography

Anderson, T., and H. Rubin. 1949. Estimators of the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics* 21: 570–582.

Andrews, D., M. Moreira, and J. Stock. 2006. Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* 74: 715–752.

Berndt, E. 1991. *The practice of econometrics, classic and contemporary*. Reading: Addison-Wesley.

Bound, J., D. Jaeger, and R. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90: 443–450.

Dagenais, M., and J.-M. Dufour. 1991. Invariance, nonlinear models, and asymptotic tests. *Econometrica* 59: 1601–1615.

Dufour, J.-M. 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65: 1365–1388.

Durbin, J. 1954. Error in variables. *Review of the International Statistical Institute* 22: 23–32.

Engle, R. 1984. Wald likelihood ratio and Lagrange multiplier tests in econometrics. In *Handbook of econometrics*, ed. Z. Griliches and M. Intriligator, Vol. 2. Amsterdam: North-Holland.

Hausman, J. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1272.

Hood, W., and T. Koopmans 1953. *Studies in econometric method*, vol. 14 of Cowles Foundation Monograph. New York: Wiley.

Horowitz, J. 2001. The bootstrap. In *Handbook of econometrics*, ed. J. Heckman and E. Leamer, Vol. 5. Amsterdam: North-Holland.

Kleibergen, F. 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70: 1781–1803.

Kleibergen, F. 2004. Expansions of GMM statistics that indicate their properties under weak and/or many instruments and the bootstrap. Working paper, Brown University.

Kleibergen, F. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73: 1103–1124.

Moreira, M. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71: 1027–1048.

Nelson, C., and R. Startz. 1990. Some further results on the exact small sample properties of the instrumental variables estimator. *Econometrica* 58: 967–976.

Nerlove, M., and F. Waugh. 1961. Advertising without supply control: Some implications of a study of the advertising of oranges. *Journal of Farm Economics* 43: 813–837.

Newey, W., and D. McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of econometrics*, ed. R. Engle and D. McFadden, Vol. 4. Amsterdam: North-Holland.

Staiger, D., and J. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65: 557–586.

Theil, H. 1953. *Estimation and simultaneous correlation in complete equation systems. Mimeo*. The Hague: Central Planning Bureau.

Wu, D.-M. 1973. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41: 733–750.

Theil, Henri (1924–2000)

Teun Kloek

Abstract

Henri Theil was a highly prolific writer of books and articles in the fields of econometric methods and applied econometrics. His best-known methodological contributions include the method of two-stage least squares, inequality coefficients (for evaluation of forecasting errors), and two measures for income inequality. Most of his other methodological contributions can be found in his *Principles of Econometrics* (1971). His best-known applied work is on economic forecasting, income inequality, applied demand analysis, and consumption patterns.

Keywords

Consumer demand systems; Cowles Commission; Distribution-free statistics; Econometric equation systems; Estimation; Florida model; Forecasting; Gricliches, Z.; Income inequality; Inequality coefficient; Information theory;

Kendall–Theil estimator; Limited information maximum likelihood method; Monte Carlo studies; Outliers; Rational random behaviour; Rotterdam model demand system; Siegel, A.; Simultaneous equations; Stabilization policy; Statistical decomposition analysis; Theil, H.; Theil–Sen estimator; Three-stage least squares; Two-stage least squares

JEL Classifications
B31

Henri Theil was a highly prolific writer of books and articles in the fields of econometrics, statistics and applied economics. The best known of his many contributions are the method of two-stage least squares, two inequality coefficients for the evaluation of forecast errors, and two measures for income inequality based on information theory, along with extensive applied work on stabilization policy, demand systems and consumption patterns.

Theil was born in Amsterdam in 1924. His parents moved first to Gorinchem and later to Utrecht, where he attended the gymnasium and obtained his diploma in 1942. He started as a student of chemistry, physics and mathematics at Utrecht University. There he was exposed to, among other subjects, formal calculus, which he disliked. In his later life he disdainfully used to call it 'epsilontics'. When he was confronted with the option between a loyalty oath to the German Nazi occupying forces and working in Germany, he decided to go underground, as did the majority of his fellow students. After some time he was caught and sent to a concentration camp, where he was incarcerated for six months. When he became seriously ill with diphtheria, his parents managed to get him out through bribery.

After the war, in 1945 he enrolled as a student at the University of Amsterdam and studied economics. In addition, he attended lectures in mathematical statistics and was a research assistant of Professor Daniel van Dantzig, at that time the most prominent Dutch statistician. As these lectures emphasized foundations rather than applications, Theil taught himself the rest of statistics by studying the two volumes of Kendall (1943, 1946). In 1951 he obtained his Ph.D. and in the same year he married Eleonore Goldschmidt. From 1951 to 1955 he worked at the Central Planning Bureau in The Hague, the main institute of policy analysis in the Netherlands, and one of the important advisers to the Dutch government.

In 1953 Theil was appointed Professor of Econometrics at the Netherlands School of Economics at Rotterdam (now Erasmus University), at first on a part-time basis while he continued his job at the Central Planning Bureau. He spent the academic year 1955–56 as a visitor at the Cowles Commission in Chicago. There he learnt the rule: 'publish or perish'. In September 1956 he obtained a full-time position in Rotterdam, where he became the first director of the Econometric Institute of the Netherlands School of Economics. This gave him the possibility to appoint research associates and assistant professors to work for him. In the ten years of his period in the Econometric Institute he attracted 27 foreign visitors, most of whom spent a year at the institute, the most prominent ones being A. S. Goldberger, A. Zellner, F. M. Fisher, M. Nerlove and Z. Griliches. In the same year Theil also started a programme in quantitative economics, where matrix algebra and mathematical statistics were prerequisites for his courses in econometric theory.

In 1965 Theil was appointed University Professor at the University of Chicago, one of the ten posts of this type at the university. In 1981 he moved to the University of Florida (at Gainesville), where he accepted the first Eminent Scholar Chair in Florida's State University System (the McKethan–Matherly Chair). He retired in 1994, but thereafter published several articles and in 1996 a book. He died in 2000.

During his life he wrote at least 18 books and about 250 published articles. Several books and about half of the articles were co-authored by more than 80 colleagues, most of them his juniors. He also supervised 15 dissertations (for details, see Raj and Koerts 1992). He could be very generous, but a smooth cooperation with him required either acceptance of his authority or outstanding diplomatic gifts. There were numerous rumours

about frictions and conflicts with colleagues and others, but written evidence is usually lacking.

Theil wrote his first well-known paper (1950), a contribution to distribution-free and robust statistics, while he was a research assistant. Consider a set of points $(x_1, y_1), \ldots, (x_n, y_n)$ and consider a pair (x_i, y_i) , (x_j, y_j) with $x_i \neq x_j$. Then the slope of the line through such a pair of points is given by $(y_j - y_i)/(x_j - x_i)$. He proposed the median of all these slopes (for $x_i < x_j$) as a distribution-free estimator of the regression line through the n points. This is an outlier robust estimator in the sense that this median is based on two 'good' points (that is, points that are not outliers) provided that the fraction of outliers is less than 0.293. In addition, it is a quite efficient estimator. The method is sometimes called the Kendall–Theil estimator, since Theil's paper made use of a result of M. G. Kendall on rank correlation. It was extended by Sen (1968) for the case that there are ties among the x observations. This method is known by the name 'Theil–Sen estimator'. Finally, Theil's approach was crucially improved by Siegel (1982), who proposed the repeated median estimator. Siegel's first step takes for each i the median across all j of the above ratios ($j \neq i$), and the second step takes the median of all medians obtained in the first step. This estimator does not even break down for a fraction of outliers close to 0.5 (see Rousseeuw and Leroy 1987, for a more extensive treatment of this subject).

Theil's best known contribution to econometrics, the two-stage least squares (2SLS) method, dates from his period at the Central Planning Bureau. He proposed it in a paper (1953) that was unpublished at the time, but it was soon known at the Cowles Commission at the University of Chicago, in those days the most important centre for econometric theory. In fact, Radner and Bobkowski (1954) of the Cowles Commission wrote a short paper in which they explained 2SLS using Cowles Commission notation. The first published description of 2SLS can be found in Theil's monograph *Economic Forecasts and Policy* (1958), while the original 1953 paper has been reprinted in Raj and Koerts (1992). The problem of systems of simultaneous equations is that most of these equations contain current (that is, nonlagged) endogenous explanatory variables. As a rule, these are not independent of the disturbances of the equation, which leads to inconsistent parameter estimates if (ordinary) least squares is applied. The 2SLS method for estimating such an equation consists of two steps. First, the current endogenous explanatory variables are regressed on the exogenous and lagged endogenous variables of the system. Second, least squares is applied to the original equation where the current explanatory endogenous variables are replaced by the explained parts of the corresponding first stage regressions. The main virtue of the method was that it was a considerable simplification on limited information maximum likelihood, the method proposed four years earlier by Anderson and Rubin (1949). This method requires the computation of eigenvalues of a matrix, a considerable burden at a time when very few universities and other research institutions had computers. (For a more extensive discussion of 2SLS and related methods, see Davidson and MacKinnon 1993.)

The Dutch Central Planning Bureau published each year a number of forecasts on several important macroeconomic variables. Theil's tasks included the evaluation of these forecasts. One of the criteria he developed for that purpose was the *inequality coefficient*, defined by

$$U_1 = \frac{\sqrt{\frac{1}{n}\sum (P_t - A_t)^2}}{\sqrt{\frac{1}{n}\sum P_t^2} + \sqrt{\frac{1}{n}\sum A_t^2}},$$

where P_t and A_t denote time series for the predicted and actual changes, respectively, of a certain variable (see Theil 1958). An attractive property of the formula is that it is positive and does not exceed unity. The zero lower bound corresponds with perfect forecasts, the unit upper bound with the worst possible set of forecasts. Unfortunately this coefficient has a serious drawback. The denominator depends on the chosen forecasting procedure, while when different forecasting procedures are to be compared it is desirable that the denominator is the same. Granger and Newbold (1973) gave a more formal

criticism using a time series model. Several years earlier Theil (1955) had already proposed an alternative formula:

$$U_2 = \frac{\sqrt{\frac{1}{n}\sum (P_t - A_t)^2}}{\sqrt{\frac{1}{n}\sum A_t^2}},$$

which is not subject to the above criticisms. It exceptionally takes values greater than unity and it is exactly equal to unity in the case of no change extrapolation ($P_t = 0$ for all t). In (1966) Theil stated why he preferred U_2 to U_1, but he used the same symbol (U) and the same term (inequality coefficient) for both U_1 and U_2 with the unfortunate consequence that if 'the' Theil inequality coefficient is cited one does not know whether the author has used U_1 or U_2 unless an explicit definition or reference is given. His criticism on U_1 was not generally noticed, as is demonstrated by the fact that U_1 survives in several places, for instance, in the well-known computer application Eviews.

Theil's most important research projects in his Rotterdam years concerned forecasting, stabilization policy, and information theory. The policy project was quite ambitious. The theory of consumer demand, where a consumer is assumed to maximize his utility function subject to his budget constraint, served as a paradigm. He assumed a government with a quadratic 'welfare function' to be maximized subject to the reduced form of a linear (or linearized) macroeconometric model, taking account of the uncertainty caused by the disturbances of the model. This implied linear decision rules for the government. While the theory was already largely developed in his 1958 book, Theil applied the theory in his 1964 book to the US economy in the 1930s and to the Dutch economy in the period 1957–9. After 1964 he left the subject to others.

Theil's book on economics and information theory appeared in 1967 but was entirely written before he left Rotterdam in September 1966. The idea came up in a discussion with Barten and the author in 1962 or 1963 on Barten's first version of the demand system that would later obtain the name 'Rotterdam model'. Barten (1964) had proposed a system of differenced demand equations. Theil amended this idea, first, by proposing premultiplication by the corresponding current budget shares w_i (before assuming constant coefficients), and later replaced the current budget shares by the averages of current and lagged budget shares, as was customary in the computation of the Törnqvist approximation to the growth rate of the Divisia quantity index. When commenting on this discussion in *Studies in Global Econometrics* Theil stated, after having mentioned Barten's research: 'In 1965 I modified this approach *slightly* . . .' (emphasis added). Not everybody was happy about the Rotterdam model, however. It was criticized at an early stage by McFadden (see Yoshihara 1969).

When Theil, in exploring the consequences of the assumptions underlying the Rotterdam model, arrived at terms of the type $\sum w_i \log w_i$, he recognized this as an expression from information theory, and started to study the literature on that subject (see also Raj and Koerts 1992, Vol. 1, pp. 25–6). Theil proceeded to develop several economic applications of information formulas. The best known is his measure for income inequality

$$\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{x_i}\right)$$

where n is the number of subsets in a population, x_i the population share of subset i in the total population, and y_i its income share in total income. The formula has attractive decomposition properties, which explains its popularity among applied researchers. In the same book he also proposed a second measure, which is obtained from the above formula by interchanging the roles of y and x. In certain situations this is considered to be preferable.

For his lectures in Chicago Theil wrote a new version of his notes on econometric methods, which appeared in 1971 under the title *Principles of Econometrics*. This book was one of the most important reference texts in the field of econometric methods during the 1970s and later.

It is also a useful reference to most of his other methodological contributions, such as his work on aggregation, on specification analysis, on the *k*-class, on mixed estimation (with A. S. Goldberger), on three-stage least squares (with A. Zellner), on the final form of econometric equation systems (with J. C. G. Boot), on efficient estimation of disturbances, and on several other subjects.

In his Chicago period, Theil also extended his work on statistical decomposition analysis (1972) and on consumer demand systems (1975, 1976). In the early 1970s he wrote three papers containing an economic theory of the second moments of disturbances of behavioural equations, also called rational random behaviour (see his 1975 and 1980 books for summaries and references). This was an ambitious project, which, however, met with little response in the profession.

The demand systems were relatively rich in parameters, with the consequence that asymptotic results were unreliable. Theil organized several Monte Carlo studies (as a rule carried out by his students or associates) in order to get more reliable test results. A survey of this work is given in Theil (1986).

In 1978 *Economics Letters*, a new journal, was launched. It specialized in short papers, only lightly refereed and published with minimum delay. Theil very much liked this formula and during the following 20 years he published (with several coauthors) a huge number of papers in the journal on a wide variety of subjects. His principal interest was now in world income inequality and international consumption patterns. For the latter purpose he developed a new type of demand system, later called the 'Florida model', and a new estimation method based on information theory. Surveys of most of this research can be found in Theil (1987, 1989, 1996).

Three of Theil's books became citation classics: *Economic Forecasts and Policy* (1958), *Economics and Information Theory* (1967), and *Principles of Econometrics* (1971). He obtained honorary degrees from the University of Chicago (1964), the Free University of Brussels (1974), Erasmus University Rotterdam (1983) and Hope College in Michigan (1985).

Given Theil's enormous productivity, the above survey of his work is necessarily very incomplete. The three volumes edited by Raj and Koerts (1992) contain considerably more information about his work up to 1992. Also of interest is a Festschrift edited by Bewley and Tran Van Hoa (1992) that appeared at about the same time. The interview by Bewley (2000) gives a good impression of Theil as a person and research worker. In all these publications one can also find more about his life (see also Kloek 2001, 2002). In his final years he developed a taste for autobiographical writing (see Theil 1996, pp. 1–6, 1997).

See Also

- ▶ Econometrics
- ▶ Forecasting
- ▶ Griliches, Zvi (1930–1999)
- ▶ Two-Stage Least Squares and the k-class Estimator
- ▶ Wage Inequality, Changes in

Selected Works

1950. A rank-invariant method of linear and polynomial regression analysis. *Proceedings of the Royal Netherlands Academy of Sciences* 53, 386–392, 521–525, 1397–1412.

1953. Ch 6. Estimation and simultaneous correlation in complete equation systems. Mimeographed Memorandum of the Central Planning Bureau, vol. 1. The Hague. Reprinted in Raj and Koerts (1992).

1955. Who forecasts best? *International Economic Papers* 5, 194–199.

1958. *Economic forecasts and policy.* Amsterdam: North-Holland.

1964. *Optimal decision rules for government and industry.* Amsterdam: North-Holland.

1966. *Applied economic forecasting*. Amsterdam: North-Holland.
1967. *Economics and information theory*. Amsterdam: North-Holland.
1971. *Principles of econometrics*. New York: Wiley.
1972. *Statistical decomposition analysis: With applications in the social and administrative sciences*. Amsterdam: North-Holland.
1975, 1976. *Theory and measurement of consumer demand*, vol. 2. Amsterdam: North-Holland.
1980. *The system-wide approach to microeconomics*. Oxford: Basil Blackwell.
1986. (With T. Taylor and J. Shonkwiler.) Monte Carlo testing in systems of equations. In *Advances in econometrics*, vol. 5, ed. D. Slottje and G. Rhodes, Jr. Greenwich: JAI Press.
1987. (With K. Clements) *Applied demand analysis: Results from system-wide approaches*. Cambridge, MA: Ballinger.
1989. (With C.-F.Chung and J. Seale, Jr.) *International evidence on consumption patterns*. Greenwich: JAI Press.
1996. *Studies in global econometrics*. Dordrecht: Kluwer.
1997. Rotterdamse Herinneringen [Rotterdam memories]. In *Kritisch en Constructief* [Critical and Constructive], ed. H. Van Dijk et al. Rotterdam: Liber Amicorum for T. Kloek, Econometric Institute, Erasmus University.

Bibliography

Anderson, T., and H. Rubin. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20: 46–63.

Barten, A. 1964. Consumer demand functions under conditions of almost additive preferences. *Econometrica* 32: 1–38.

Bewley, R. 2000. Mr Henri Theil: An interview with the International Journal of Forecasting. *International Journal of Forecasting* 16: 1–16.

Bewley, R., and Tran Van Hoa. 1992. *Contributions to consumer demand and econometrics*. London: Macmillan.

Davidson, R., and J. MacKinnon. 1993. *Estimation and inference in econometrics*. New York: Oxford University Press.

Granger, C., and P. Newbold. 1973. Some comments on the evaluation of economic forecasts. *Applied Economics* 5: 35–47.

Kendall, M. 1943, 1946. *The advanced theory of statistics*, vol. 2. London: Griffin.

Kloek, T. 2001. Obituary: Henri Theil, 1924–2000. *Statistica Neerlandica* 55: 263–269.

Kloek, T. 2002. *Henri Theil in Rotterdam, 1953–1966*. Amsterdam: Paper presented at the Theil Memorial Conference.

Radner, R., and Bobkoski, F. 1954. Ch 8. Some recent work of H. Theil on estimation in systems of simultaneous equations. Unpublished discussion paper of the Cowles Commission, vol. 1. Reprinted in Raj and Koerts (1992).

Raj, B., and J. Koerts. 1992. *Henri Theil's Contributions to Economics and Econometrics*. Vol. 3 vols. Dordrecht: Kluwer.

Rousseeuw, P., and A. Leroy. 1987. *Robust regression and outlier detection*. New York: Wiley.

Sen, P. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63: 1379–1389.

Siegel, A. 1982. Robust regression using repeated medians. *Biometrika* 69: 242–244.

Yoshihara, K. 1969. Demand functions: An application to the Japanese expenditure pattern. *Econometrica* 37: 257–274.

Theory Appraisal

Ellery Eells and Daniel M. Hausman

Abstract

Economists and other scientists appraise theories in terms of criteria such as evidential support, predictive accuracy, usefulness, and reliability in research and practice. This entry addresses three general problems concerning theory appraisal. 1. What is it for evidence to be confirmationally relevant to a theory? 2. How can evidential support be measured? 3. On what other criteria should theory appraisal in economics depend? The third question raises special problems, since economic models so often incorporate statements that appear to be false. But answers to general questions concerning confirmation matter to the conduct of economics, too.

Keywords

Bayes, T.; Bayesian confirmation theory; Ceteris paribus; Confirmation; Duhem–Quine problem; Experimental economics; Friedman, M.; Habit; Hempel, C.; Hume, D.; Hypothetico-deductive method; Induction; Likelihood; McCloskey, D.; Mill, J. S.; Popper, K.; Prior probability; Probability; Rhetoric of economics; Subjective probability; Testing; Theory appraisal

JEL Classifications

B4

There are several reasons why economists and others appraise theories and models. They may want to judge whether theories and models are worthy of study, whether one can rely on them in research and practice, or whether one can judge them to be true or false or predictively adequate. Different purposes may call for different appraisals.

All sciences face three general problems of theory appraisal. 1. What is it for evidence e to confirm, disconfirm, or be evidentially irrelevant to an hypothesis *h*, and how is confirmation possible? 2. How should scientists measure the extent to which *e* confirms *h* or how well confirmed *h* is overall? 3. How does the appraisal of a theory depend on the extent to which it is confirmed, and on what else may theory appraisal depend? The third question raises special problems with respect to economics, because the conclusions of economics are usually hard to test, whether experimentally or against market data. Economic models incorporate many statements that appear to be false, and the basic principles of economics typically contain, at least implicitly, vague and ineliminable *ceteris paribus* qualifications. These peculiarities are not unique to economics, though their combination and the particular form that questions of theory appraisal take in economics are distinctive.

Answers to general questions concerning confirmation and theory appraisal have immediate implications for the conduct of economics. For example, defenders of real business cycle theory have constructed computational 'experiments' in which technology shocks in a calibrated computer model of a simplified economy give rise to simulated business cycles. Are these appropriate methods for testing economic theories? Do the results of these 'experiments' justify a positive appraisal of real business cycle theory (Kydland and Prescott 1996)?

The discussion of confirmation and its measurement, which occupies the next two sections, is quite general, while the discussion of theory appraisal in the final section emphasizes distinctive features of economics.

Confirmation

Evidence *e* confirms an hypothesis *h* when it provides some support for *h*; *e* disconfirms *h* when it provides some evidence against *h*. The theory of confirmation is not concerned with the limiting cases of conclusive proof or refutation. Unlike deductive inferences, in which it is impossible for the premises to be true and the conclusion false, confirmation is a matter of inductive inference, which is fallible. From the two premises, 'All humans are mortal' and 'Bill Gates is human', one can deduce that Bill Gates is mortal. Though infallible, a deductive inference such as this one is also 'non-ampliative' – the conclusion is already implicit in the premises, and drawing the conclusion arguably does not advance our knowledge. To infer that Bill Gates will die from the premises that he is human and that all humans born before 1850 have died is, in contrast, to take a risk. The conclusion goes beyond what is asserted in the premises. The conclusion that *all* humans are mortal obviously goes far beyond any data concerning past births and deaths.

David Hume (1748) issued a serious challenge to the rationality of inductive inferences and to the idea that there is such a thing as confirmation or rational support. As an empiricist, he insisted that all evidence derives from perception and thus can be stated as reports of observations. He then asked what argument employing only observation reports as premises could establish universal

generalizations or conclusions concerning phenomena not yet observed. Such arguments could not be deductive, since they are fallible. There was nothing logically impossible about Europeans observing black swans in Australia, even though all swans Europeans previously observed had been white. To get from premises reporting observations to a generalization or a claim about something not yet observed requires as an additional premise or as a rule of inference some principle to the effect that nature is, in the relevant respect, uniform or regular. But Hume points out that we are entitled to rely on such a premise only if it is a logical truth or is itself established by experience. A principle of uniformity is not a logical truth, and it is no easier to establish on the basis of reports of past uniformities than the particular induction we are attempting to justify. Hume draws the extreme sceptical conclusion that humans have no good reason at all to believe the conclusions of inductive inferences, or, in other words, that observational evidence never provides any rational support for any hypotheses. What keeps people out of harm's way and enables them to survive is habit. Observing regularities, people cannot help expecting them to persist.

Hume confesses that, once he leaves his study, he cannot take his sceptical conclusions seriously. Furthermore, treating confirmation as no more than a matter of conditioned responses to repetition leaves one with few ways to distinguish between sensible uses of evidence on the one hand and prejudice, superstition or even phobia on the other. Hume's problem of induction shows that the notion of justification that is relevant to science and everyday life is piecemeal rather than global. In inductive arguments, scientists can legitimately employ premises that are not observation reports and are hence not beyond questioning. (And, of course, observation reports are fallible, too.) In order to clarify and inform scientific practice, theories of confirmation need to capture the distinction between evidence that supports an hypothesis and evidence that does not, rather than to deny that this distinction exists.

In thinking about confirmation, it is helpful to draw three distinctions. First, one may distinguish between 'incremental' confirmation – confirmation as a relation between an hypothesis h and a particular piece of evidence e – and total or absolute confirmation – as a judgement about how well supported h is overall or as a relation between h and the total available evidence. Evidence e confirms h incrementally if e provides (or would provide) some additional support for h, whether or not h is well confirmed in the sense of total confirmation. Second, confirmation can be thought of either *qualitatively* or *quantitatively*. In the total sense, h is either qualitatively well confirmed or poorly confirmed, while quantitative confirmation theory attempts to quantify how well or poorly confirmed h is. In the incremental sense, evidence e may, qualitatively, either confirm, disconfirm, or be evidentially irrelevant to an hypothesis h, while quantitative theory attempts to measure how much e boosts or diminishes how well supported h is. Finally, confirmation can be either *comparative* or *non-comparative*. Non-comparative confirmation concerns just one hypothesis–evidence pair. In comparative confirmation, on the other hand, one assesses, for example, how well an e supports an h relative to how well a different e' supports the same h or how well the same e supports a different h'.

A simple and natural idea is that a general hypothesis of the form 'All Fs are Gs' is (incrementally) confirmed by 'positive instances' – that is, by objects that are both F and G. A negative instance – an object that is an F but not a G – does not merely disconfirm the general hypothesis; negative instances refute universal generalizations. This somewhat misleading asymmetry plays a large role in Karl Popper's insistence that science relies on confirmation rather than verification. For an example of instance confirmation, an increase in a firm's sales following a drop in the price of its product confirms the law of demand. If in addition to instance confirmation one assumes that logically equivalent hypotheses are confirmed by the same evidence, one is immediately faced with a paradoxical conclusion. A white shoe confirms the generalization 'Everything that is not black is not a raven', which is logically equivalent to the generalization 'All ravens are black'. But do reports of white shoes confirm 'All ravens are black' (Hempel 1945)?

The most common of the many responses to this paradox rely on a quantitative theory of confirmation and are discussed briefly below.

The positive instance criterion was offered as part of a sufficient condition for the confirmation of universal conditionals. Hempel (1945) generalized this criterion to his *satisfaction criterion* which applies to more complex logical structures for evidence and hypothesis, and which provides explicit definitions of confirmation, disconfirmation, and evidential irrelevance. In defending his satisfaction criterion, Hempel relied on what he regarded as intuitively obvious *conditions of adequacy* for definitions of confirmation. Besides the *equivalence condition* mentioned above, the *entailment condition* covers the limiting case where evidence logically entails an hypothesis: it says that, if e entails h, then e confirms h. Hempel also assumed a more controversial *special consequence condition*, which says that, if e confirms h and h entails h', then e confirms h'.

The criteria for confirmation discussed above apply when evidence reports and hypotheses are stated in the same language, which Hempel took to be an observational language. What about confirmation of *theories*, which typically contain theoretical as well as observational vocabulary? *Hypothetico-deductivism* (HD) or the *hypothetico-deductive* method is the idea that theories and hypotheses are confirmed by their observational consequences. So, for example, axioms concerning individual preference are confirmed by the same evidence that supports the law of demand, even though that evidence is not an instance of the preference axioms, because that evidence (and the law of demand itself) can be deduced from models which include axioms governing individual preferences. The positive instance and satisfaction criteria are formulations of the idea, roughly, that observations that are *logically consistent with* a hypothesis confirm the hypothesis, while HD says that *deductive consequences of* a theory or model confirm the theory or model. Of course, if the prediction fails (if an observational deductive consequence of a theory turns out to be false), then this is supposed to provide *disconfirmation* of the theory.

The HD method is subject to two serious problems. The first is the problem of irrelevant conjunction or of distributing credit. If an hypothesis h logically implies an observation report e, then so does the conjunction, $h\&g$, where g can be any sentence whatsoever. If, as the HD method says, an observation report e confirms any hypothesis that implies e, then e confirms $h\&g$. That seems bad enough. Worse still, since $h\&g$ implies g, the special consequence condition implies that e confirms g, or in other words any sentence whatsoever. A natural response would be to refine the basic HD idea and deny that e confirms h when h entails e, unless some further condition is met to the effect that there is no logically weaker hypothesis h' that also entails e. But these difficulties are not easily remedied.

The second problem, which is particularly poignant in economics, concerns the distribution of blame. Testable implications can rarely if ever be deduced from scientific theories without additional premises. Deductive logic tells us that, if those testable implications are false, then at least one of the premises from which they follow must be false, but deductive logic alone does not single out which is the culprit. This is the so-called 'Duhem–Quine problem' and is discussed later.

Quantitative Confirmation: Probabilistic Approaches

The most influential probabilistic approach to confirmation is called 'Bayesian confirmation theory'. The basic idea is that evidence e confirms hypothesis h if and only if the conditional probability $p(h/e)$ (defined as $p(h\&e)/p(e)$) is greater than the unconditional probability $p(h)$. Disconfirmation is defined by reversing the inequality, and evidential irrelevance is defined by changing the inequality to an equality. The function p measures an agent's 'subjective probability' or degree of belief. Given a set of axioms governing preference, which are stronger than those that must be satisfied in order to attribute ordinal utility functions to agents, one can prove that there exists a 'cardinal' utility representation, which is unique up to a positive affine transformation, and

which assigns to each lottery a utility equal to the sum of the utilities of its prizes weighted by the agent's subjective probabilities. Such cardinal representation theorems (for example, Harsanyi 1977, ch. 4) establish that the degrees of belief of agents who satisfy the axioms conform to the axioms of the probability calculus. (Important work in foundations of subjective probability includes Ramsey 1931; de Finetti 1937; Savage 1972; Jeffrey 1983; Joyce 1999.)

Suppose that $p(h/e)$, the agent's subjective conditional probability, is equal to what the agent's subjective degree of belief in h would be if the agent learned that e is true – that is, in the wake on an observation that e is true, the agent changes or updates $p(h)$ to be equal to the current $p(h/e)$. On this assumption $p(h/e) > p(h)$ if and only if the agent's degree of belief in h would increase if e were learned. In that case, it is natural to say that, for this agent, e is positively evidentially relevant to h, even when e is not in fact learned. So the Bayesian maintains that e confirms h if and only if $p(h/e) > p(h)$, that e disconfirms h if and only if the inequality is reversed, and e is evidentially irrelevant if and only if $p(h/e) = p(h)$.

People's subjective probabilities will differ, even if they are equally rational, because their background knowledge differs. So, according to the Bayesian approach, confirmation is a relation among three things: evidence, hypothesis and background knowledge. The reason this approach is called 'Bayesian' confirmation theory is because of its frequent use of a mathematical theorem discovered by Thomas Bayes (1764), a simple version of which is: $p(h/e) = p(e/h)p(h)/p(e)$. This expression follows easily from the definition of conditional probability which tells us that $p(h/e) = p(h\&e)/p(e)$ and $p(e/h) = p(h\&e)/p(h)$. Bayes's theorem shows how the quantity of interest, $p(h/e)$, depends on the so-called 'prior probability' – $p(h)$ – the 'likelihood' – $p(h/e)$ – and the subjective probability of the evidence. When $p(e)$ is low – that is, when an observation is surprising – then one has much stronger evidence than if the observation was expected.

Bayesian confirmation theory also suggests *measures* of the *degree* of evidential support that e provides for h. The most common measure is the *difference measure*: $d(h, e) = p(h/e) - p(h)$, where confirmation, disconfirmation and evidential irrelevance correspond to whether this measure is positive, negative or zero, and degree goes by the magnitude of the difference, but there are other measures, too.

One application of the idea of degree of evidential support has been to the ravens paradox, discussed above. Let h be the hypothesis that all ravens are black; let Ra symbolize the statement that object a is a raven; and let Ba symbolize the statement that a is black. If h is probabilistically independent of Ra (that is, $p(h/Ra) = p(h)$), then a positive instance (or report of one), $Ra\&Ba$, of h confirms h in the Bayesian sense (that is, $p(h/Ra\&Ba) > p(h)$) if and only if $p(Ba/Ra) < 1$ (that is, if it is not *already* certain that a would be black if a raven). Given the same independence assumption, one can prove that a positive instance, $Ra\&Ba$, confirms h more than a contrapositive instance, *not-Ba&not-Ra*, does, provided that one's degree of belief that a given raven will be black is less than one's degree of belief that something that isn't black isn't a raven. Since for most people the subjective probability that a particularly non-black thing such as a white shoe is not raven is much higher than the subjective probability that any particular raven is black, black ravens will provide much stronger confirmation of 'All ravens are black' than white shoes or pale economists.

Bayesian confirmation theory can also be used to assess Hempel's proposed conditions of adequacy for criteria of confirmation. In particular, Hempel's special consequence condition (unlike the equivalence conditions or the entailment condition) does not follow from the Bayesian definition of confirmation and the axioms of the probability calculus, and by attending to the circumstances in which the special consequence condition can fail, theorists have constructed intuitively compelling examples of an h entailing an h', and an e confirming h while disconfirming h'. Such examples tell against the special consequence condition and also in favour of Bayesian confirmation theory (for example, Eells 1982).

One standard objection to Bayesian confirmation theory is that, if e is already known to be true,

then $p(h/e)$ must be the same as $p(h)$ and, by the Bayesian definition of confirmation, e cannot confirm h. This is called 'the problem of old evidence'. One possible Bayesian solution to the problem, suggested by Glymour (1980), would be to say that it is not the already known e that confirms h, but rather a newly discovered logical or explanatory relation between e and the particular h. Other solutions have been proposed and various versions of the problem have been distinguished (see Earman 1992, for discussion). The problem remains one of lively debate.

A second objection to Bayesian confirmation theory is more general and leads to the main contemporary probabilistic alternative. Prior probabilities – the degree of belief agents have in hypotheses in advance of gathering particular evidence – and likelihoods – $p(h/e)$ – play crucial roles in Bayesian confirmation theory. But is it plausible to suppose that these are known? If h is a newly formulated physical hypothesis, for example, what would *justify* an assignment of probability to it prior to evidence? One response to the problem is to point to convergence theorems (as in de Finetti 1937) which show that differences in priors will 'wash out' in the long run. But this response is not very satisfactory, since assessments of theories need to be made now, and if initial differences in priors are large enough, the long run must be long indeed.

Those who favour a *likelihood* approach to the evaluation of evidence – for example, Edwards (1972) and Royall (1997) – believe that there are important contexts in which likelihoods can be known, but that reliance on prior probabilities is rarely justifiable. According to one formulation of the likelihood account, e confirms h more than e confirms h' if and only if $p(h/e)$ is greater than $p(e/h')$. This is a comparative principle. It doesn't matter that both of these likelihoods may be tiny. The likelihood approach separates the question of which hypothesis it is more justified to believe given all the evidence from the question of which hypothesis is better supported by the particular piece of evidence. There are a variety of different measures of confirmation in terms of likelihoods (see Fitelson 2001, and Forster and Sober 2002, for recent discussion and references). Although those who emphasize likelihoods avoid relying on priors, they cannot of course avoid relying on likelihoods, and these too may be hard to know. Although evidence statements are often entailed by the conjunction of the hypothesis under test and a variety of other statements concerning the specific circumstances, measuring apparatus, the absence of interferences and so forth, the conditional probability of the evidence statement conditional on the particular hypothesis may be as hard to nail down as the prior probability.

Inexactness and Theory Appraisal

The appraisal of theories ought presumably to be strongly influenced by how well confirmed they are both absolutely and in comparison with alternatives. If there is very strong evidence in support of a theory – that is, if a theory is very well confirmed – then the risk of believing it and relying on it both in practice and for the purposes of gaining further knowledge should be small.

But, given how ambitious and wide-reaching the claims of theories often are, it is questionable whether even the best-supported of scientific theories is all that well supported. How could evidence from our little corner of the universe justify conclusions about all bodies and all forces? Furthermore, other features of theories should influence appraisal of them, in addition to how well supported they are. Some theories are more 'promising' than others. Some are easier to use or to teach. Some theories are more compatible than others with strongly held metaphysical and religious beliefs or with other branches of science. A fruitful falsehood may be worth more than a tedious truth. Appraising theories and choosing among them depends on much more than how well confirmed they are.

Appraising economic theories is particularly challenging because of the limited possibilities for experimentation and because of the limited relevance of market data, which are influenced by many factors from which economic theories abstract. The Duhem–Quine problem is a huge practical problem rather than just a philosophical possibility. Consider, for example, Card and

Kreuger's (1995) study of the effect of an increase in minimum wages in New Jersey in 1990, which found, in contradiction to the prediction of simple economic models, that unemployment among unskilled workers did not decline. To carry out this test, they compared employment in fast-food restaurants in New Jersey, where minimum wages increased, and next-door Pennsylvania, where minimum wages did not change. Their results might show, as they maintained, that relatively small changes in the minimum wage do not cause unemployment among unskilled workers. But the results might instead be due to peculiarities of fast-food restaurants, to other changes in the economy or popular culture of New Jersey or Pennsylvania, or to flaws in their data. The problem is ubiquitous: market data bear on economic theories through so many intermediary premises that successes of predictions give economists little reason for confidence in their models, and failures give economists little reason for dissatisfaction with their models. To bridge the gap between even the very best data and economic theories requires a motley assortment of simplifications and *ceteris paribus* clauses, which rarely inspire much confidence. When predictions fail, it is typically more sensible to blame the problem on the simplifications or *ceteris paribus* conditions than on the theory. That means that economists can hang on to their basic principles without much risk of refuting them but also without much prospect of improving them.

If one turns from predictions given by economic models to the propositions out of which the models are constructed, one finds that they rely heavily on premises such as 'Commodities are infinitely divisible', 'All agents have perfect information concerning prices and quantities of all commodities', or 'Consumption decisions are taken by a single representative consumer'. No one thinks these claims are true, but many believe that their falsity does not matter. Perhaps they can be regarded as approximations. Perhaps they can be regarded as idealizations. Perhaps they can be regarded as abstracting from details that are, for some purposes, irrelevant. In addition, many of the basic principles or 'laws' of economics – claims such as 'Agents' preferences are transitive' or 'Consumers prefer more commodities to fewer' – are false, at least if they are construed as universal generalizations. Yet they seem somehow to capture important facts that enable economists to build models that predict and explain important economic phenomena.

The traditional view of theory appraisal in economics receives its best exposition in the works of John Stuart Mill (1843, Book VI). Although an empiricist, Mill thought in terms of causation. Individual causes can be and ought to be studied by observation (if one is fortunate enough to find a domain where no other causal factors are interfering) or experimentally (if one can construct controlled circumstances which differ only with respect to the presence or absence of the particular cause one investigates). Mill's famous methods of induction guide inferences concerning individual causes (1843, Book III).

When one is studying phenomena that are influenced by many causes, like the phenomena that economists study, Mill maintains that direct inductive methods cannot be used. Instead one needs first to determine the laws of the main causes by studying other domains where their action can be isolated, and then one needs to deduce their effect when combined. Mill believes that it is possible to learn the basic behavioural generalizations governing economic behaviour by means of introspective psychology on the one hand and engineering on the other. Economics then takes the form of exploring the implications of models in which economists combine what they know of the most important causal factors with simplified descriptions of the circumstances to which the model is to be applied. So, for example, economists know that firms generally aim to maximize net revenue and that they have a choice of different production processes which employ relatively more or relatively less unskilled labour. When minimum wage laws increase, wages of unskilled labour increase (unless market rates are already higher than the new minimum wage), and firms will economize on the use of unskilled labour. Employment of unskilled labour will consequently decline. Mill urges in addition that empirical tests such as the Kreuger and Card study be carried out as checks on economists'

deductions and to determine whether they have left out some significant factor. But predictive failures do not show any errors in one's basic principles, which have, in Mill's view, already been conclusively established by the results of introspective observation of the principles separately. The appraisal of basic theory depends on the quasiexperimental study of its individual basic 'laws' or principles, and it is largely disconnected from the appraisal of specific models, which depends on whether they serve the particular purposes for which they are constructed.

Although Mill's view that the basic principles of economics are proven truths is unsustainable, he was right to emphasize the difficulty of testing economics against uncontrolled market data and to stress the value of evidence concerning specific principles gathered from everyday experience. As the example of the employment effects of minimum-wage laws illustrates, economists rely on everyday experience to suggest and justify generalizations concerning the behaviour of firms including the flexibility of their employment policies. Causal experience cannot, of course, show how single-minded or relentless firms are in maximizing net returns, or whether this causal factor is one of a small set of factors that is of supreme importance with respect to virtually all economic phenomena. But, given the complexities and ambiguities in empirical studies such as Card and Krueger's, it is not unreasonable to continue to expect increases in unemployment among unskilled workers to result from increases in minimum wages, and it is not unreasonable to hang on to fundamental principles such as profit maximization.

In the second half of the 20th century, many economists became uneasy about espousing a view of theory appraisal that emphasized the extent to which the basic principles of economics were independently confirmed or disconfirmed, often by casual experience. One reason was that empirical work apparently showed that firms do not attempt to maximize profits (Hall and Hitch 1939; Lester 1946). Rather than addressing the details of these studies, Milton Friedman (1953) argues famously that such inquiries into the 'realism' of the 'assumptions' of economics reflected methodological confusion. All that matters is how successfully economic models are at predicting the market price and quantity data that are of interest to economists. Friedman's view appeared to economists to be in better accord with an up-to-date philosophy of science. It was empiricist, and, later on, many saw it as corresponding roughly to Karl Popper's insistence on falsification (Blaug 1976, p. 149). Given how difficult it is to test economic theory against market data, Friedman's view of theory appraisal in fact served to insulate economics from empirical criticism, which may have been its main appeal.

During the second half of the 20th century methodologists flirted with a number of views of theory appraisal, many of which were constructed with the natural sciences or even mathematics in mind, and none of which has found general favour among economists or economic methodologists. The difficulties of theory appraisal have even led some influential voices to argue for an abandonment of the pursuit of any normative standards. McCloskey (1985) thus apparently argues that, apart from very unspecific requirements of honesty and open discussion, there are no rules constraining the appraisal economists offer of their theories, while Hands (2001) suggests that the only fruitful questions to ask concerning theory appraisal are questions about the sociological factors that lead communities of scientists to endorse one theory or another.

In our view, in contrast, it is impossible to evade normative questions about how theories and models ought to be appraised. Though these questions have no simple answers, the answers must turn in large part on matters of confirmation. Decisions to rely on particular claims of economics both for theoretical and for practical purposes are weighty indeed, and it would be massively irresponsible to base policy on some economic thesis without asking how well confirmed it is.

See Also

- ▶ Assumptions Controversy
- ▶ Ceteris Paribus
- ▶ Experimental Economics

▶ Falsificationism
▶ Mill, John Stuart (1806–1873)
▶ Models
▶ Philosophy and Economics
▶ Rhetoric of Economics
▶ Science, Economics of
▶ Testing

Bibliography

Bayes, T. 1764. An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London* 53: 370–418.
Blaug, M. 1976. Kuhn versus Lakatos *or* Paradigms versus research programmes in the history of economics. In *Method and appraisal in economics*, ed. S. Latsis. Cambridge: Cambridge University Press.
Card, D., and A. Krueger. 1995. *Myth and measurement: The new economics of the minimum wage*. Princeton: Princeton University Press.
de Finetti, B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincaré* 7, 1–68. English translation in *Studies in Subjective Probability*, ed. H. Kyburg and H. Smokler. New York: John Wiley, 1964.
Duhem, P. 1954. *The aim and structure of physical theory*, trans. P. Wiener. Princeton: Princeton University Press, 1954.
Earman, J. 1992. *Bayes or bust: A critical examination of bayesian confirmation theory*. Cambridge, MA/London: MIT Press.
Edwards, A. 1972. *Likelihood*. Cambridge: Cambridge University Press.
Eells, E. 1982. *Rational decision and causality*. Cambridge/New York: Cambridge University Press.
Fitelson, B. 2001. Studies in Bayesian confirmation theory. Ph.D. dissertation, University of Wisconsin-Madison.
Forster, M., and E. Sober. 2002. Why likelihood? In *The nature of scientific evidence*, ed. M. Taper and S. Lee. Chicago: University of Chicago Press.
Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.
Glymour, C. 1980. *Theory and evidence*. Princeton: Princeton University Press.
Hall, R., and C. Hitch. 1939. Price theory and business behaviour. *Oxford Economic Papers* 2: 12–45.
Hands, D. 2001. *Reflection without rules: Economic methodology and contemporary science theory*. Cambridge: Cambridge University Press.
Harsanyi, J. 1977. *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.
Hempel, C. 1945. Studies in the logic of confirmation. *Mind* 54, 1–26, 97–121. Reprinted, with his Postscript, in *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: Free Press, 1965.
Hume, D. 1748. *An inquiry concerning human understanding*, 1955. Indianapolis: Bobbs-Merrill.
Jeffrey, R. 1983. *The logic of decision*. 2nd ed. Chicago/London: University of Chicago Press.
Joyce, J. 1999. *The foundations of causal decision theory*. Cambridge/New York: Cambridge University Press.
Kydland, F., and E. Prescott. 1996. The computational experiment: An econometric tool. *Journal of Economic Perspectives* 10(1): 69–85.
Lester, R. 1946. Shortcomings of marginal analysis for wage-employment problems. *American Economic Review* 36: 62–82.
McCloskey, D. 1985. *The rhetoric of economics*. Madison: University of Wisconsin Press.
Mill, J.S. 1843. *A system of logic*, 1949. London: Longman, Green & Co..
Ramsey, F. 1931. Truth and probability. In *The foundations of mathematics and other logical essays*, ed. R. Braithwaite. London: Routledge and Kegan Paul.
Royall, R. 1997. *Statistical evidence: A likelihood paradigm*. Boca Raton: Chapman and Hall.
Savage, L. 1972. *The foundations of statistics*. 2nd ed. New York: Dover Publications, Inc..

Theory of Economic Integration: A Review

Nigel Grimwade

Abstract

The theory of economic integration is the branch of economics concerned with analysing the effects of different forms of integration on the economies of member states and the rest of the world. Its relevance for Europe is the progress made since the foundation of the European Community and European Free Trade Area in 1958 and 1960 in dismantling trade barriers, adopting a common external tariff (in the case of the EC), establishing a single market and, more recently, creating a common currency. The basic theory of customs union, first expounded by Viner in 1950 and later extended by Meade and Lipsey, provides the theoretical foundation on which the theory

of integration rests. While Viner's work was important in showing that customs unions and free trade areas are not always welfare-enhancing and may even lower global economic welfare, in its simple form the theory was incomplete. It focused mainly on the short-run effects of regional integration and failed to provide a convincing rationale for why countries enter into such arrangements. Subsequently, Viner's analysis was modified and added to by relaxing some of the more limiting assumptions on which it rested, preparing the way for a deeper understanding of the integration process. In particular, our understanding of how integration affects countries was strengthened by the incorporation of economies of scale and terms of trade effects, which Viner had largely ignored.

Beginning in the 1980s, important advances were made by extending the analysis to incorporate the effects of increasing returns and imperfect competition. An important role in this respect was played by the emergence of the new trade theories. The launching of the Single Market programme in 1987 led to greater attention being given to the effects of deep integration on markets in which intra-industry trade was the predominant form of competition. On top of the normal gains from lower prices and improved resource allocation, potentially much greater gains could be reaped from intra-industry specialisation. At the same time, integration theory became much more interested in the effects of integration on economic growth. The application of endogenous growth theories to integration theory appeared to show that much the largest gain from integration results from a permanent increase in the regional growth rate. More recently, integration theory has become concerned about the location effects of integration, reflecting the growing interest of trade theorists in the importance of geography. New models of trade, incorporating the effects of factor mobility, external economies of scale and product competition, have established the importance of location in the analysis of the effects of integration. In short, integration theory has come a long way from where it started out fifty years or more ago, leaving us with a much more comprehensive picture of how it impacts on countries both inside and outside the region.

Keywords

Customs unions; Dynamic effects; Economic growth; Economic welfare; Economies of scale; Free trade areas; General equilibrium; Imperfect competition; Location; Partial equilibrium; Single market; Static effects; Terms of trade

JEL Classifications

F15; F42; F43; F55

Introduction

The term 'economic integration' is used to describe a process whereby the economies of several different countries, often in close geographical proximity, are bound together into a single region. This may happen through a process of governments reducing barriers to the free movement of goods, services, persons or capital (negative integration) or through governments creating common policies and institutions for the purpose of regulation and control (positive integration). The terms 'deep' and 'shallow' integration have also been used to distinguish measures applied at the border (e.g. tariffs and quantitative restrictions) from measures that are domestic or 'behind the border' (e.g. harmonisation, service liberalisation, investment liberalisation, competition policy). However, the term may also be used to the describe the *outcome* of this process in the form of increased intra-regional trade, greater price convergence, increased specialisation, increased economic interdependency, an increase in intra-regional capital flows, a reduction in interest-rate spreads, etc. Although economic integration may occur between countries at a global level, the focus of this article is on the regional dimension as it is encountered in different types of regional trading agreements – free

trade areas, customs unions, common markets and economic and monetary unions.

The European Context

In Europe, economic integration began with the establishment of a customs union of six countries (West Germany, France, Italy and the Benelux countries) following the signing of the Treaty of Rome in March, 1957 and the creation of a free trade area of seven countries (the UK, Sweden, Denmark, Norway, Austria, Switzerland and Portugal) two years later following the signing of the Stockholm Convention. In 1987, following the passage of the Single European Act, the nine members of the European Community (EC) embarked on the internal market programme designed to bring about the establishment of a true single market by 1992. The latter was defined as a market based on the so-called 'four freedoms' – free movement of goods, services, persons and capital. This was followed by the creation of a European Economic Area (EEA) in which the privileges and obligations of the single market (but excluding agriculture) were extended to the members of EFTA, excluding Switzerland. (Austria, Sweden and Finland subsequently became full members of the EC.) The move towards monetary union had to await the passage of the Treaty on European Union (the Maastricht Treaty) in 1991, which was signed by the 15 members of the former EC. This committed the member states to move towards full monetary union, including the adoption of a common currency, single central bank and common monetary policy by 1997 subject to enough member states satisfying the required degree of convergence. Under a special derogation, the UK was not required to participate in this process except following a separate vote in favour of doing so of the UK House of Parliament. This exemption was also subsequently extended to Sweden and Denmark. The final stage of the transitional arrangements for achieving full monetary union commenced on 1 January, 1999 leading eventually to the adoption of the euro as the common currency of the Eurozone, initially by 11 of the 15 member states. Following enlargement of the EU in 2005, membership of the Eurozone increased to 17 countries. Thus, Europe has exhibited elements of all forms of integration, involving both the removal of barriers and other forms of discrimination and the creation of new regional institutions and policies. These have included measures applied at the border and a 'behind the border' dimension. The outcome has been apparent in an increase in the share of intra-EU trade in the total trade of the member states, increased intensity of intra-EU trade, increased (vertical and horizontal) specialisation and some further evidence of policy and price convergence in the Single Market (see de Lombaerde 2006).

The Basic Theory of Customs Unions

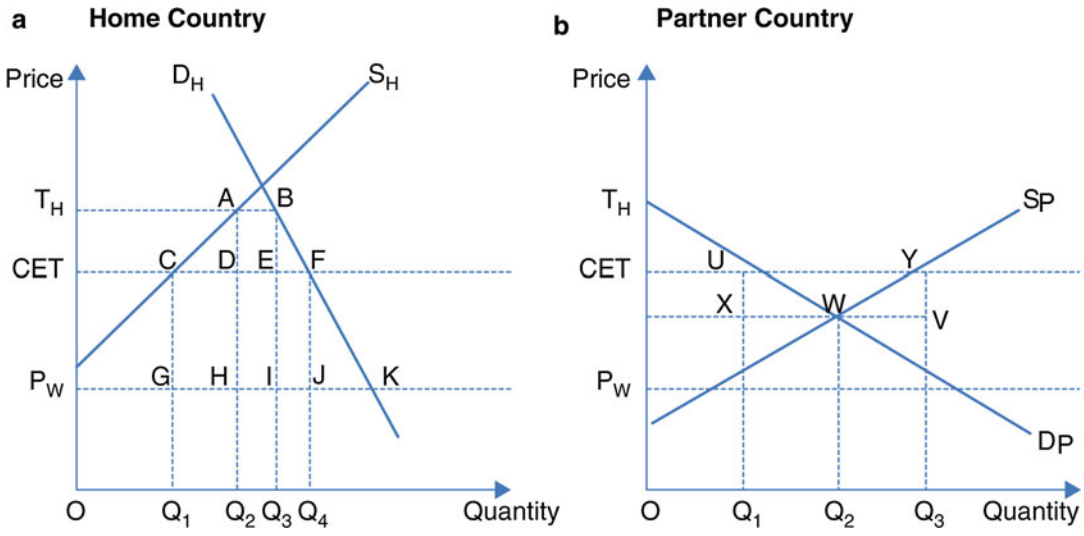
The first attempt to develop a theory of regional integration was primarily concerned with the static effects of customs unions on resource allocation. In this regard, the most important contribution was unquestionably the work of Viner (1950), who challenged the then prevalent view that customs unions constitute steps towards free trade because they lower trade barriers and thus increase trade. He demonstrated that whether or not a customs union raises or lowers global economic welfare cannot be determined *a priori*. This is because customs unions have two effects on resource allocation, one of which raises economic welfare while the other reduces it. First, the removal of barriers to trade between the member states brings about an increase in intra-union trade as production shifts from high-cost to low-cost producers within the union. Such *trade creation* results in a better allocation of resources benefitting both countries, which are now able to reallocate resources towards economic activities in which they are relatively efficient. However, because customs unions are discriminatory, favouring imports from member states over imports from nonmember states, they also result in *trade diversion*. The reduction in tariffs on imports from member states means that some goods that were previously imported from a low-cost source outside the union are now

imported from a higher-cost source within. As production shifts from a low-cost to high-cost source, the outcome is a misallocation of resources in the global economy and thus a fall in global economic welfare. Thus, while trade creation takes the world closer towards free trade, trade diversion takes the world closer towards protectionism. The net effect on economic welfare of the formation of a customs union thus depends on the relative strength of these two effects. These will depend on several factors, including the size of the customs union, the level of the common external tariff (relative to the average tariff of the member states before the formation of the union), the degree of complementarity or rivalry between the members of the union and the differences between members in the unit costs of their protected industries of the same kind (i.e. their potential complementarity). Net trade-creating unions are beneficial, while net trade-diverting unions are harmful.

The importance of Viner's work was in showing that the establishment of a customs union (or free trade area) is not necessarily welfare-improving viewed from a global economic perspective. However, as was subsequently pointed out by Meade (1955), Viner's analysis is incomplete as it considers only the *production* effects of a customs union – that is, the efficiency gain in production from the re-allocation of resources. The formation of a customs union, however, will also have *consumption* effects resulting from a change in the relative prices of different goods following integration. If we assume that most goods have an elasticity of demand greater than zero, the removal of tariffs will lower the price of importable goods leading to an increase in demand for the goods and increased consumption. This creates a source of gain for consumers, who can buy more of these goods than they could before and pay less for each, resulting in an increase in their consumer surplus. Account must also be taken of the loss of tariff revenue to member states when trade expansion takes the form of trade diversion. Inclusion of the consumption effects, however, increases the potential welfare gain to countries from the formation of customs union. As Lipsey (1957) demonstrated, if favourable consumption effects outweigh unfavourable production effects, even a trade-diverting customs union could be welfare-improving both for the member states and the world.

These effects are generally illustrated using a partial equilibrium framework in a model consisting of two countries (home and partner) and the rest of the world. Goods are assumed to be homogenous and produced under conditions of increasing marginal costs (i.e. upward-sloping supply curve) and perfect competition. This is shown in Fig. 1.

Before the formation of the customs union, given a world, free-trade price of O-P_W, the home country imposes a tariff of P_W-T_H with the country consuming O-Q_3, producing O-Q_2 and importing Q_2-Q_3. However, the partner country, which is a lower-cost producer of the good, imposes a tariff of P_W-T_P such that the country consumes and produces O-Q_2 but imports none of the good. Assume the two countries form a customs union and adopt a common external tariff of P_W-CET which is a weighted average of the pre-union tariffs applied by both countries. In the home country, the price falls from O-T_H to O-CET, consumption increases to O-Q_4, production falls to O-Q_1 and imports increase to Q_1-Q_4. In the partner country, the price rises from O-T_P to O-CET, consumption falls to O-Q_1, production increases to O-Q_3 and the country exports Q_1-Q_3. The home country experiences both trade-creation and trade-diversion effects. The amount of trade created is given by the increase in imports of Q1-Q4 less Q2-Q3 which results in a net welfare gain to the importing country given by the triangles ADC and BFE. (Conceptually, this is the difference between the gain to consumers from increased surplus CET-T_H-B-F and the loss to producers of producer surplus given by CET-T_H-A-C and loss of tariff revenue of A-B-E-D.) However, because the CET is a discriminatory tariff, imports now come from the partner country and not the rest of the world, so resulting in trade diversion. The loss to the importing country from trade diversion is given by the difference between the free-trade price O-P_W and the cost of imports from the partner country P_W-CET and



Theory of Economic Integration: A Review, Fig. 1 The effects of a customs union on home and partner countries

the imports diverted Q2-Q3 which is the area DEIH. The gain to the partner country from trade diversion comes in the form of increased producer surplus T_P-CET-Y-W less the loss of consumer surplus from higher prices given by CET-U-W-T_P. This is the area UYW.

A further assumption is that costs are increasing, thus ruling out any effects of integration from market enlargement resulting in economies of scale. If, however, average costs fall with output and the country faces high tariffs in foreign markets, they may be unable to reach a scale of production large enough to exploit all available economies of scale. Viner was aware that a customs union could create a home market in which firms could grow to a sufficient size to exploit such economies. However, he doubted whether the size of the market was a constraint except for small countries in all but a few industries. It was left to Corden (1972) to provide a formal analysis of the effects of the formation of customs unions where average costs decrease with output. He showed that, in a situation where a good is produced in both countries before the formation of the union, integration has both a conventional trade creation effect *and* a 'cost reduction effect'. The latter is an additional gain from the formation of the customs union that results from a cheapening of existing sources of supply rather than a movement to cheaper sources. A further possibility is that the formation of the union induces production of a good in one of the two countries which previously produced none of the good at all. This would happen if the latter had lower production costs than the established producer in the other country, but previously imported the good from the rest of the world. Corden termed this the 'trade suppression effect' to distinguish it from trade diversion proper.

The major weakness of the Viner theory of customs unions is that it rests too heavily on overly restrictive assumptions. In addition to those already stated, it is assumed that the union is too small to have any effect on the world price of imports (i.e. it faces a perfectly elastic world supply curve). It is also assumed that the common external tariff is set at the average of the level of tariffs of the member states before the union was formed. Firms are assumed to produce under conditions of perfect competition. There are also no externalities, so that resources are optimally allocated where profit-maximising firms equate price to marginal cost. The framework used is also a partial equilibrium one that ignores any feedback effects between countries or sectors. Subsequent

attempts to extend the theory of regional integration have sought to address these deficiencies.

The Rationale for Customs Unions

One criticism of the Viner theory is that it lacks any economic rationale for why countries form customs unions. Cooper and Massell (1975) showed that the formation of a net trade-creating customs union is always 'second best' to an equivalent, non-preferential reduction in tariffs. Whereas a non-discriminatory cut in tariffs leads to trade creation only, a customs union will usually result in both trade creation and trade diversion. In Fig. 1, a non-preferential tariff cut by the home country of CET-T_H would have an identical effect on production and consumption as forming a customs union with the partner country. Both result in the price of the good in the home country falling to O-CET and lead to imports increasing by the same amount. In a customs union, however, the home country suffers a loss of tariff revenue given by ABIH, because all imports now come from the partner country. With a non-preferential cut, tariff revenues may rise or fall (being given by the difference between CFJG less ABIH), but will always leave revenues higher than the case of a customs union. Thus the importing country is always better off making a non-preferential tariff cut, except, of course, in the case where the partner country is the lowest cost supplier of the good in the world.

Why, then, do countries form customs unions? One reason is that countries consider there to be possible *export advantages* from joining a customs union. Wonnacott and Wonnacott (1981) criticised Cooper and Massell for paying too much attention to the interests of the importing country. The orthodox theory makes the implicit assumption that the home country faces no external tariff on the goods that it exports to the rest of the world. This follows from the assumption which it makes that the customs union is small, in which case other countries may see little benefit from imposing tariffs on exports from the union. The Wonnacotts referred to this as the problem of the 'missing' third-country tariff. It is more reasonable to assume, they argued, that even small countries face tariffs on their exports because tariffs provide countries with a potential bargaining counter that can be used to negotiate better access for exports. If, then, a country is a low-cost producer of a good, it may be prevented from fully exploiting its comparative advantage because of high tariffs in the rest of the world. By forming a customs union and adopting a common external tariff, however, the country may succeed in expanding its output of these goods and so boost its economic welfare. Of course, it could have achieved the same effect by making a unilateral tariff reduction rather than forming a customs union. This, however, assumes that the rest of the world will follow suit by unilaterally lowering its own tariffs, which they may not.

In the search for a rationale for customs unions, Cooper and Massell (1965) argued that the motives for countries forming customs unions may be non-economic. Specifically, customs unions might be established as a less costly way of protecting markets where governments have non-economic policy objectives. Johnson (1965) criticised the standard theory for using a welfare function that included only *private* consumption of goods by individual households and ignored the collective consumption of goods that satisfy *public* wants. An important reason for governments of countries at a low stage of economic development imposing tariffs is to meet a collective preference of consumers for industrial over non-industrial goods. Although tariffs impose a welfare loss, such action is justifiable where the social benefits from increased collective consumption of industrial goods exceed the private costs of protection. For the same reason, governments of countries that are net exporters of industrial goods may grant an export subsidy supported by an import tariff to prevent re-importation. In practice, however, governments are often prevented from taking such measures by international agreements that prohibit the use of trade-distorting subsidies. They are therefore unable to fully satisfy the collective preference of consumers for these goods. Customs unions, however, provide a means whereby they can do so through partner countries reciprocally reducing

their tariffs on industrial goods from each other while adopting a common external tariff on these goods coming from the rest of the world. If all the members of the union have a strong collective preference for industrial production, they can, thereby, together expand their exports of these goods without suffering any loss to their own production. Although each country will suffer a loss of consumer surplus as consumers are forced to pay more for goods supplied by the partner instead of importing from the rest of the world, producers will enjoy increased producer surplus and public demand for increased collective consumption of industrial goods will be satisfied. In this case, even a customs union that is net trade-diverting can raise economic welfare for the members concerned. As an argument to explain why countries form customs unions, however, its application would seem to be limited to countries yet to satisfy their collective wants for industrial goods, i.e. countries at a low stage of economic development. There is also a presumption that governments are unable to use industrial subsidies to increase production, since subsidies are less trade-distorting than tariffs.

Terms of Trade Effects

Other economic reasons why countries might form customs unions are evident if the assumption that the union is too small to influence the price at which goods are supplied by the rest of the world is relaxed. If, instead, we assume that a customs union faces an upward-sloping, rest-of-the-world supply curve, the possibility exists that through the formation of the union the members could improve their terms of trade. Viner was aware of this possibility, but rightly dismissed it as irrelevant to the issue of whether or not the formation of a union raises or lowers global economic welfare. This, of course, is because the terms-of-trade gain for the union is matched by an equal loss for the rest of the world. Mundell (1964) and Arndt (1968), however, later provided a formal analysis showing how customs unions influence the terms of trade. This happens because trade diversion reduces imports from the rest of the world, leading to a fall in the price of world exports and an improvement in the terms of trade of the customs union. If the effect is large enough, the members of the union could in theory be better off than before. This may not, however, be true for all the members, in which case a conflict of interest will arise. This will require the country that gains to persuade the country that loses to act in a way that is contrary to the latter's own self-interest. A further difficulty is the possibility that the rest of the world might retaliate by imposing tariffs on the exports of the members of the union. The formation of a customs union may, however, serve a further purpose, as Viner observed, namely, to strengthen the bargaining power of the member states in multilateral trade negotiations.

The terms of trade effects of customs unions have also been considered using a general equilibrium framework. The partial equilibrium framework used in the standard theory has limitations where a customs union is large and results in big changes in tariffs affecting different goods. A general equilibrium approach may also be more appropriate for determining the circumstances in which a customs union is beneficial to the member countries. Building on the work of Vanek (1965), Kemp (1969), Ohyama (1972), and Kemp and Wan (1976) showed that, if the common external tariff is set at a certain level and lump sums are transferred between members, a customs union can be established which results in trade creation only. The common external tariff should be set at a level that leaves unchanged the member countries' trade with the rest of the world in terms of both quantities and proportions. In this case, the rest of the world is not affected at all while the member states benefit from trade creation. If the external tariff were set at a lower level, the rest of the world would also gain. If all customs unions were to meet these conditions, the creation of more customs unions would be welfare-improving and take the world closer towards free trade. Customs unions could be enlarged by adding new members in such a way as to improve global economic welfare without harming the rest of the world. The significance of this is in showing that, in theory at least, customs unions have the potential to be welfare-improving.

Dynamic Effects

A major weakness of the standard theory is that it focuses entirely on the static effects of integration. These may be defined as the short-term, once-and-for-all effects of market integration, which are concerned with the effects of resource re-allocation. Empirical studies have found that, where integration leads to net trade creation, the welfare gain expressed as a percentage of GNP is relatively small. According to Mayes (1978), the estimated welfare gain from the formation of the European Community, although positive, amounted to less than 1 per cent of GNP. By way of contrast, early attempts to estimate the dynamic effects of the formation of the EC found these to be much larger (e.g. Owen 1983). The dynamic effects may be defined as the long-run, ongoing effects of integration, which result in a faster rate of economic growth. These arise from the ability of firms to produce at a point further down their long-run average cost curve due to the enlargement of the market in which they operate (including dynamic economies of scale from learning curve effects); the effects of economic efficiency resulting from the exposure of firms to increased competition (including the elimination of X-inefficiency caused by managerial slack); the boost to investment brought about by enlargement of the market and increased flows of inward direct investment; and the stimulus to innovation resulting from both the enlargement of the market and greater competitive pressures.

One reason for the greater importance attached to these dynamic effects was a growing awareness of the importance of intra-industry trade (IIT) in the trade of developed countries and, in particular, the members of the European Community (e.g. Balassa 1967, 1974; Kreinin 1979; Grubel and Lloyd 1975; Greenway and Hine 1991). As much as two-thirds of the trade in manufactured goods between developed countries was estimated to take the form of two-way trade in similar goods (goods belonging to the same industry). Much of this trade took place in horizontally and vertically differentiated goods produced in imperfectly competitive markets under conditions of decreasing costs (increasing returns). The increased importance of IIT was the result of a growing necessity for firms to spread the heavy fixed costs associated with the production of such goods over a larger output so as to reduce average cost and meet consumers' wish for increased choice and greater variety. Such intra-industry specialisation was made possible by a lowering of tariff barriers and the consequent enlargement of the market in which they could sell their goods. Empirical studies found that the gains from the formation of the European Community were much larger if account was taken of these effects. Owen (1985) estimated that the dynamic gains raised the GDP of the EC by between 3% and 6%.

Statistical evidence that much of the expansion of trade in manufactured goods between developed countries took the form of intra-industry trade led in the 1980s and 1990s to the development of new trade theories based on imperfect competition. The latter may take the form of either oligopolistic markets dominated by a few sellers of an identical product or monopolistically competitive markets with many sellers of differentiated goods. A model of trade predicated on the existence of only two firms located in different countries identical in all respects and selling a single homogenous good was developed by Brander (1981) and Brander and Krugman (1983) to show how and why intra-industry trade will result. If firms are assumed to engage in Cournot-type behaviour (taking the output decisions of their rival as given) a point of equilibrium (Nash equilibrium) before and after trade can be determined. The incorporation of transport costs leads to the prediction that each firm will sell half its output in the home market and the other half in the market of the other country. One consequence of this is the phenomenon of reciprocal dumping, in which the *ex-factory* price of the exported product is below that of the product sold at home. The major gain from such trade results from an increase in competition, which results in lower prices and the elimination of X-inefficiency due to managerial slack.

Early attempts to model trade in markets characterised by monopolistic competition mainly focused on the effects of trade on consumer choice and their preference for variety. Lancaster (1980)

developed a model of trade between countries with identical factor endowments in horizontally differentiated goods, all of which possess core attributes. Because of decreasing costs, firms are unable to produce all the varieties that consumers want in the absence of trade. By expanding the potential market in which firms can sell their goods, trade leads to more varieties of the good being produced, such that consumers taken as a whole can buy a variety closer to their preferred one. At the same time, producers can produce each variety in larger quantity and so produce at lower cost, leading to lower prices. Building on the work of Dixit and Stiglitz (1977), Krugman (1979) and Venables (1984) developed a similar model in which consumers gain from having a greater variety of the same good (so-called 'love of variety') as well as from lower prices. Subsequently, Helpman (1981) proposed a model of trade between two countries with different factor endowments capable of giving rise simultaneously to both inter- and intra-industry trade. IIT takes place between the two countries in differentiated manufactures, while inter-industry trade based on factor endowments results in both manufactures and food. Helpman and Krugman (1985) showed that, where the two countries differ in size, the larger country is a net exporter of differentiated goods.

Ethier and Horn (1984) argued for four extensions to the conventional theory of customs unions – trade modification (resulting from the removal of tariffs on goods coming from partner countries that are not imported at all from the rest of the world), small tariff changes (or marginal tariff adjustments), scale economies and product differentiation under imperfect competition. They developed a three-country model each with two industries – agriculture (produced under conditions of constant returns to scale) and manufacturing (producing differentiated goods under conditions of increasing returns to scale). The two partner countries export manufactures to each other as well as to the rest of the world, but import food from the rest of the world only. Given these assumptions, the formation of customs unions between the two partners cannot result in any trade diversion, but only trade creation and trade modification. The model is then used to examine two changes in the commercial policy of the union – an increase in the common external tariff on imports of food and the imposition of a marginal internal barrier on the exchange of differentiated manufactures. An increase in the common external tariff results in increased food production within the union, diverting resources away from manufacturing and leading to a reduction in the number of varieties available. However, because the terms of trade move in favour of the customs union, the outcome is ambiguous and dependent on other variables. The imposition of a marginal internal barrier in an attempt by each member state to protect its manufacturing again results in a decline in the number of varieties, but an increase in the production of food. In other words, the attempt to protect manufacturing surprisingly has the opposite effect of protecting agriculture. Once again, the effect on the welfare on both member states and the non-member country could go either way.

New trade theories have been applied to the measurement of the effects of European integration. The contribution of Smith and Venables (1988) was of major importance in the assessment of the effects of the EC's single market programme. Baldwin and Wyplosz (2004) adopt a framework for assessing the effects of market integration in Europe based on similar assumptions. Under imperfect competition, integration has two effects – an initial increase in competition as the number of firms increase in national markets and a restructuring process leading eventually to a decline in the number of firms. Firms, however, become larger, which enables them to produce at lower average costs. Gains accrue to countries in the form of lower prices for consumers and fall in average costs due to economies of scale.

Growth Effects

Spurred on, in part, by the opportunities created by the launching of the EC's Internal Market Programme, integration theory became

interested in the effects of integration on economic growth. Of particular importance in this regard was the work of Baldwin (1989) whose predictions of the effects of the Single Market placed the greatest emphasis on the medium- and long-run growth effects of the removal of internal barriers. The major source of increased growth in the medium term was the increase in physical capital formation induced by the integration process. Integration improves the efficiency with which factors are combined in the production process, causing an increase in output. With a constant ratio of savings to output, the rise in output induces higher savings, which in turn leads to higher investment. An increase in investment then leads to a further increase in output in a cumulative fashion. However, in a neoclassical model of economic growth, such an increase in economic growth cannot last. With a constant amount of labour, the ratio of capital to labour rises, which leads, in turn, to a fall in the marginal productivity of capital (i.e. diminishing returns), thus bringing to an end the investment boom. Eventually, the growth rate falls back to its 'steady state'.

This, however, is only true if we assume no improvement in technological efficiency. In the neoclassical theory of growth, technological progress is treated as an *exogenous* variable such that an increase in the rate of physical capital formation has no effect on total factor productivity. New growth theory, however, *endogenises* the growth process, making possible a more rapid rate of growth in the long term, i.e. a permanent change in the growth rate leading to 'ceaseless accumulation'. In endogenous growth models, investment yields increasing returns due to the gap or wedge arising between the private and public return on capital. Although the return on capital as perceived by the individual investor diminishes with the level of investment, the public return does not. The explanation for this proposed by Romer (1983, 1986) emphasised the role played by 'knowledge spillovers'. Investment undertaken by an individual firm creates technology spillovers benefitting other producers, which therefore also engage in further investment. The combined action of many firms undertaking investment boosts the social return on capital even though the investment by one firm in isolation results in diminishing returns. Subsequently, Lucas (1988) attached greater importance to the role of knowledge capital by switching the emphasis in growth models from physical to human capital.

Location Effects

A further aspect of economic integration that has more recently attracted attention in theoretical work is concerned with the effects of integration on the geographical location of economic activity. This has been influenced by the growing importance attached by trade theorists to the role played by geography in shaping the pattern of specialisation and trade between countries. Beginning with Krugman (1991), economists have developed models of trade between countries which seek to predict the influence of global and regional integration on the geographical distribution of economic activity between regions. Such models incorporate two aspects of trade that are largely ignored in the conventional approach – namely, the occurrence of increasing returns (decreasing costs) in certain activities and the presence of trade costs arising from the need to ship goods over long distances. These give rise to two conflicting forces: on the one hand, the presence of increasing returns favours the *concentration* of production in a single location and, on the other hand, trade costs promote geographical *dispersal* of activity over many locations. In these models, an important role is played by factor mobility. As trade costs fall, it becomes worthwhile for firms in peripheral regions to relocate activity to core regions. Such relocation increases competition for existing firms in the core regions, which reduces the profits of firms. However, the rise in wage rates attracts more workers to the region. This increases the demand for goods in the core region and eases competition in the labour market, increasing the profits of firms in the core region and so attracting more firms to the region. The result is a widening gap between core and peripheral regions.

An added feature of these models is the recognition that part of manufacturing is concerned with producing intermediate goods which enter as inputs in the production of other manufactured goods. Important advantages accrue to firms from producing intermediate goods close to the final goods producers. These take the form of backwards linkages (from downstream firms buying goods from suppliers) and forwards linkages (from intermediate producers supplying goods to downstream activities). Firms which concentrate their activities in the same regions, therefore, benefit from significant economies of agglomeration. Krugman and Venables (1995) examined the agglomeration effects of a lowering of trade costs in a two-country, two-sector model in which labour is immobile. One sector produces food under conditions of perfect competition, while the other produces manufactures under conditions of monopolistic competition. As trade costs fall, it becomes possible for producers in one country to supply consumers in the other through trade rather than local production, so exports commence. At a certain point, however, agglomeration effects become dominant, causing firms in one country to relocate their production close to producers in the other. Agglomeration benefits from access to a larger market (a backward linkage creating increased demand for intermediate goods) and proximity to suppliers (a forwards linkage, offering lower-cost intermediate goods) offset differences in wage rates to reinforce this process. However, as trade costs fall further, differences in factor prices outweigh linkages as the determinant of location and manufacturing moves back to the region with lower wage rates.

Puga (1988) examined the effects of integration on location that bring together the two previous models. The model assumes that labour is free to migrate across regions and between sectors and that input–output linkages from firms locating in the same region generate cost savings. Given the assumptions of the model, perfect symmetry exists between the two regions at high trade costs with equal shares of industrial output in the two regions. However, as trade costs fall, agglomeration becomes sustainable, such that, with further reductions, the symmetry between the two regions breaks. At this point, the two regions 'endogenously differentiate' into one industrialised and another de-industrialised region. If there were no labour migration across regions, however, agglomeration would happen more slowly. This is because, as the number of firms begins to increase in one of the two regions, the absence of labour mobility drives up local wages as firms can only recruit more workers from the agricultural sector. This acts to discourage firms from clustering together, although this is partially offset by the advantages that come from being close to other firms. As trade costs fall further, the cost savings from concentration decline while the wage gap remains. At a critical value of trade costs, it becomes more profitable for firms to relocate in the de-industrialised region to enjoy lower wage costs and import inputs. As they do so, the incentives for further relocation increase until a point of symmetrical equilibrium is restored between the two regions.

Conclusion

The theory of integration has come a long way from where it started in the aftermath of the Second World War. The early attempts to examine the effects of customs unions and free trade areas were concerned primarily with resource allocation effects and their implications for global economic welfare. While providing a useful framework for the analysis of these effects, the standard theory had little positive to say about regional integration. More importantly, it failed to provide a convincing rationale for why countries engage in such activities. On reflection, the problem was the strictness of the assumptions on which it rested, which had only a limited application for the contemporary world. Most of the subsequent extensions of the theory have involved the gradual relaxation of these assumptions to take into account the complexity of the integration process in practice. These have served to demonstrate that, under certain conditions, regional integration can be beneficial to the countries involved without harming the rest of the world. Indeed, to the extent

that integration has a positive impact on growth, the rest of the world may share in the gains. In a world of differentiated goods and falling average costs, these gains may also be significantly greater than they appear if attention is limited to the static effects only.

See Also

- ▶ Comparative advantage
- ▶ European Union (EU) Trade Policy
- ▶ European Union Single Market: design and development
- ▶ European Union Single Market: economic impact
- ▶ North American Free Trade Agreement (NAFTA)
- ▶ Regional and preferential trade agreements

Bibliography

Arndt, S.W. 1968. On discriminatory versus non-preferential tariff policies. *Economic Journal* 78: 971–978.

Balassa, B. 1967. Trade-creation and trade-diversion in the European common market. *Economic Journal* 77: 1–21.

Balassa, B. 1974. Trade creation and trade diversion in the European common market. *Manchester School of Economics and Social Studies* 42(2): 93–135.

Baldwin, R. 1989. The growth effects of 1992. *Economic Policy* 9: 247–282.

Baldwin, R., and C. Wyplosz. 2006. *The economics of European integration*. 2 nd ed. Maidenhead: McGraw-Hill Education.

Brander, J. 1981. Intra-industry trade in identical commodities. *Journal of International Economics* 11: 1–14.

Brander, J., and P. Krugman. 1983. A reciprocal dumping model of international trade. *Journal of International Economics* 13: 313–321.

Cooper, C., and D. Massell. 1965. Towards a general theory of customs unions in developing countries. *Journal of Political Economy* 75: 724–727.

Corden, W.M. 1972. Economies of scale and customs union theory. *Journal of Political Economy* 80: 465–475.

De Lombaerde, P., ed. 2006. *Assessment and measurement of regional integration*. London: Routledge.

Dixit, A.K., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Ethier, W., and H. Horn. 1984. A new look at economic integration. In *Monopolistic competition and international trade*, ed. H. Kierzkowski. Oxford: Clarendon Press.

Greenaway, D., and R. Hine. 1991. Intra-industry specialisation, trade expansion and adjustment in the European economic space. *Journal of Common Market Studies* XXIX: 6.

Grubel, H.G., and P.J. Lloyd. 1975. *Intra-industry trade: the theory and measurement of international trade in differentiated goods*. London: Macmillan.

Helpman, E. 1981. International trade in the presence of product differentiation, economies of scale and monopolistic competition. *Journal of International Economics* 11: 305–340.

Helpman, E., and P. Krugman. 1985. *Market structure and foreign trade: increasing returns, imperfect competition and the international economy*. Cambridge, MA: MIT Press.

Johnson, H.G. 1965. An economic theory of protectionism, tariff bargaining and the formation of customs unions. *Journal of Political Economy* 73: 256–283.

Kemp, M. 1969. *A contribution to general equilibrium theory of preferential trading*. Amsterdam: North-Holland.

Kemp, M., and H. Wan. 1976. An elementary proposition concerning the formation of customs unions. *Economic Journal* 6: 95–97.

Kreinin, M. 1979. *The Effects of European Integration on Trade Flows in Manufactures*. Seminar Paper No. 125, Institute for International Economic Studies, Stockholm University.

Krugman, P. 1979. Increasing returns, monopolistic competition and international trade. *Journal of International Economics* 9: 469–479.

Krugman, P. 1991. *Geography and trade*. Cambridge, MA: MIT Press.

Krugman, P.R., and A.J. Venables. 1995. Globalisation and the inequality of nations. *Quarterly Journal of Economics* 110(4): 857–880.

Lancaster, K. 1980. Intra-industry trade under perfect monopolistic competition. *Journal of International Economics* 10: 151–175.

Lipsey, R.G. 1957. The theory of customs unions: trade-diversion and welfare. *Economica* 24: 40–46.

Lucas, R. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.

Mayes, D.G. 1978. The effects of economic integration on trade. *Journal of Common Market Studies* 17: 1–25.

Meade, J.E. 1955. *The theory of customs unions*. Amsterdam: North-Holland.

Mundell, R.A. 1964. Tariff preferences and the terms of trade. *Manchester School of Economics and Social Studies* 32: 1–13.

Ohyama, M. 1972. Trade and welfare in general equilibrium. *Keio Economic Studies* 9: 37–73.

Owen, N. 1983. *Economies of scale, competitiveness and trade patterns within the European community*. Oxford: Oxford University Press.

Puga, D. 1988. The rise and fall of regional inequalities. *European Economic Review* 43: 303–334.
Romer, P. 1983. Dynamic competitive equilibria with externalities, increasing returns and unbounded growth. *PhD Thesis*, University of Chicago.
Romer, P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1037.
Smith, A., and A.J. Venables. 1988. 1988: Completing the internal market in the EC: Some industry simulations. *European Economic Review* 32: 1501–1525.
Vanek, J. 1965. *General equilibrium of international discrimination*. Cambridge, MA: Harvard University Press.
Venables, A.J. 1984. Multiple equilibria in the theory of international trade with monopolistically competitive industries. *Journal of International Economics* 16: 103–121.
Viner, J. 1950. *The customs union issue*. New York: Carnegie Endowment for International Peace.
Wonnacot, P., and R. Wonnacot. 1981. Is unilateral tariff reduction preferable to a customs union? The curious case of the missing foreign tariffs. *American Economic Review* 71: 704–714.

Thin Markets

Marzena Rostek and Marek Weretka

Abstract

A thin market is a market with few buying or selling offers. The concept of market thinness, while general, is typically used in the context of financial markets. When the number of buying or selling offers is small, investors' trading positions are large relative to market size. Trading then requires price concessions and thus exerts an impact on prices. A thin market is characterized by low trading volume, high volatility and high bid–ask spreads. This article discusses the modelling of thin markets, some typical phenomena of such markets, and their implications for market design.

Keywords

Asset pricing; Blockage discount; Inventory models; Liquidity; Market efficiency; Market power; Market structure; Oligopoly; Overshooting; Predatory trading; Thin markets

JEL Classifications

D43; D53; G11; G12; G14; L13

A thin market is a market with few buying or selling offers. It is also known as a narrow market. The signature characteristic of a thin market is traders' price impact. When the number of buying or selling offers is small, investors' trading positions are large relative to market size. Trading then requires price concessions and thus exerts an impact on prices. A thin market is characterized by low trading volume, high volatility, and high bid–ask spreads. The concept of market thinness, while general, is typically used in the context of financial markets.

Market thinness is a particular source of market illiquidity. Liquidity is broadly defined as the ease of trading a security. Apart from market power, lack of liquidity can result from asymmetric information, transaction costs, search and bargaining frictions, imperfect financing ability, limited commitment and spatial considerations.

Price Impact in Financial Markets

Since transaction-level data became available in the early 1990s, it has been well understood that institutional investors (such as mutual funds, hedge funds, pension funds and investment banks), whose trade on the New York Stock Exchange (NYSE) accounts for more than 70 per cent of daily trading volume, exert a significant impact on prices and take this into account in their trading strategies. The seminal empirical studies include Chan and Lakonishok (1993, 1995), and Keim and Madhavan (1995, 1996, 1998). To mitigate the adverse effects of price impact, large traders do not place their orders at once; rather, they break them up into smaller blocks, which are then placed sequentially. For instance, at the NYSE, only about 20 per cent of the total trading value of all institutional purchases and sales is completed within a single trading day, while more than 50 per cent takes at least four days for execution. If traded at once, a typical institutional package would represent more than 60 per cent of the total trading

volume. In financial slang, institutional investors are referred to as elephant traders and institutional trading blocks as iceberg orders. In fact, the trading costs associated with such price impact dominate the explicit costs of trade, such as commission, order processing and brokerage fees. Consequently, extensive resources are devoted to estimating price impact and designing best execution. Such techniques are available to institutional as well as retail investors in the form of software called market impact models.

Modelling Thin Markets

On the theory side, the presence of market power poses challenges to modelling. In particular, the competitive approach is not suitable for modelling thin markets since it assumes that no individual trader can affect the market price by their buying or selling orders.

A large body of literature has emerged to explain how price impact affects individual portfolio choices of investors and the equilibrium in financial markets. The theoretical mechanisms underlying these models can be grouped into three categories: asymmetric information, inventory effects and nonequilibrium mechanisms. Traditionally, the leading class of models with price impacts is based on asymmetric or private information (e.g. Glosten and Milgrom 1985; Kyle 1985, 1989; Easley and O'Hara 1987; Back 1992; Foster and Viswanathan 1996; Holden and Subrahmanyam 1996). In such models, price impact arises because high sales by an informed trader are interpreted by remaining traders as a signal of low asset value, and hence reduce the asset price.

For many market events that involve anticipated demand or supply shocks, such as preannounced inclusions of new stock into the S&P, the price impact component that derives from asymmetric information can only partially explain the observed magnitudes of price changes. Therefore, an alternative strand of the literature based on inventory effects has emerged. There, because of diversification concerns, risk-averse traders are willing to absorb large, risky orders only at price concessions (Ho and Stoll 1981; Grossman and Miller 1988; Vayanos 2001; Attari et al. 2005; Brunnermeier and Pedersen 2005; Pritsker 2005; DeMarzo and Urošević 2006 extended by Urošević 2005). These papers capture price impact by building Cournot-type models with one or several large investors and a continuum of (infinitesimally) small price-taking traders.

Under Cournot market structure, large investors trade only with small competitive traders. In contrast, it is a stylized fact of financial markets that large investors trade effectively with one another in the sense that an order placed by a large investor is primarily absorbed by (a possibly small group of) other large traders. Essentially, this is a market structure of bilateral oligopoly, except that all traders can buy and sell. The seminal papers embedding these features are Kyle (1989) and Vayanos (1999). An equilibrium model of symmetric-information market environments in which all buyers and sellers have price impact can be found in Weretka (2006). Rostek and Weretka (2007) study dynamic thin markets and establish implications of market thinness for asset pricing. Rostek and Weretka (2008) examine the implications of market thinness for information aggregation, efficiency and price discovery. Carvajal and Weretka (2007) show that the pricing kernel exists in thin markets.

Finally to study such markets, several nonequilibrium models with price impact have been proposed (e.g. Bertsimas and Lo 1998; Almgren and Chriss 2000; Subramanian and Jarrow 2001; Dubil 2002; Almgren 2003; Huberman and Stanzl 2004; Almgren et al. 2005; Engle and Ferstenberg 2007). These models assume motivated empirically functional forms of price impact functions, one for every trader, which are then used to analyse market dynamics.

Thin Market Phenomena

Handling large orders through order break-up is but one difference between thin and competitive markets. Market thinness leads to a number of other empirical phenomena that are hard to reconcile with competitive equilibrium asset pricing models, such as CAPM of C-CAPM.

Pareto Inefficiency

Because of the reduction of buying and selling orders in response to market power, traders do not fully diversify the idiosyncratic risk of their holdings. As a result, allocations are not efficient. This is optimal, since the benefits from diversification need to be balanced against the extra (with respect to the competitive, price-taking, setting) cost of price impact.

Response to Liquidity Shocks

As evidenced by the empirical literature, the exogenous shocks in asset supply, such as inclusions of new stocks into the S&P, weight changes in stock market indices, or forced liquidations, result in price overshooting. Even if the shock is preannounced, on the date of the actual event, the price drops below the new fundamental value to attain that value only in subsequent periods. This phenomenon, often referred to as price overshooting, cannot occur in the competitive model, where prices adjust to the new fundamental value immediately following the shock or its announcement. The overshooting effect is the equilibrium reaction of thin markets to anticipated and unanticipated shocks in asset supply. The overall observed price change can be decomposed into temporary and permanent components. The permanent effect, which occurs upon the announcement, represents the adjustment of the fundamental value that results from the changes in inventory. The effect is amplified by temporary price concessions demanded by liquidity providers to be willing to absorb the shock on the day of its occurrence whether or not the shock is anticipated. An alternative explanation of overshooting involves predatory trading (Brunnermeier and Pedersen 2005): When a large trader needs to liquidate a portfolio quickly, other investors sell and subsequently buy back the asset. This strategy lowers the price at which they can obtain the liquidated portfolio.

Excess Volatility and Volatility Clustering

One of the consequences of price overshooting in thin markets is excess price volatility. That is, the presence of the price impact leads to excess return volatility and changes in volatility unrelated to changes in fundamentals. Since, in addition, the price impact varies over time, periods of high price impact feature high price volatility, thereby inducing volatility clustering.

Limits to Arbitrage

Another novel feature of thin markets is the coexistence of anticipated price differentials and limits to arbitrage in equilibrium. According to the competitive theory of asset pricing, whenever there are anticipated price differentials, a trader can make infinite profit by taking unbounded positions. When a market is thin, however, price impact naturally limits the benefits from arbitrage for active traders and also reduces incentives to enter the market. Therefore, unlike in a competitive model, the profits from entering the market are bounded, and even small fixed entry costs may prevent outsiders from arbitraging the price differentials. These entry costs include explicit trading costs, such as transaction costs, but also the cost associated with learning the characteristics of the stocks. Empirically, it may take months for outside capital to bid prices back to their fundamental value (Mitchell et al. 2007).

Asset Valuation

In the presence of price impact, the market value of a large block of shares no longer coincides with the cash value that could be obtained by liquidating the portfolio. To account for the difference, valuation specialists often apply a so-called blockage discount. A typical instance where blockage discounts are applied involves the transfer of a property in a case of divorce. It is in the interest of the divorcees to claim a large price impact (and blockage discount) which implies a large tax discount. The practical approach is based on the implementation shortfall (Perold 1988), which measures the difference between the closing or arrival price and the final execution price.

Implications for Market Design

Market thinness has further prompted changes in market design towards automation of the trade execution. To ease competition through the trading

cost of price impact, many exchanges have adopted an electronic trading system with posted orders (e.g. Nasdaq, NYSE, Euronext, and the stock exchanges in London, Toronto and Vancouver). In the presence of asymmetric information, market thinness can perversely reduce liquidity under continuous trading. Therefore, several markets have returned to more traditional trading systems.

See Also

- ▶ Arbitrage
- ▶ Countervailing Power
- ▶ Efficient Markets Hypothesis
- ▶ Exchange
- ▶ Liquidity Constraints
- ▶ Pareto Efficiency

Bibliography

Almgren, R. 2003. Optimal execution with non-linear impact functions and trading enhanced risk. *Applied Mathematical Finance* 10: 1–18.

Almgren, R., and N. Chriss. 2000. Optimal execution of portfolio transactions. *Journal of Risk* 3: 5–39.

Almgren, R., C. Thum, E. Hauptmann, and H. Li. 2005. Equity market impact. *Risk* 18: 57–62.

Attari, M., A. Mello, and M. Ruckes. 2005. Arbitraging arbitrageurs. *Journal of Finance* 60(5): 2471–2511.

Back, K. 1992. Insider trading in continuous time. *Review of Financial Studies* 5(3): 387–409.

Bertsimas, D., and A. Lo. 1998. Optimal control of execution costs. *Journal of Financial Markets* 1: 1–50.

Brunnermeier, M., and L. Pedersen. 2005. Predatory trading. *Journal of Finance* 4: 1825–1863.

Carvajal, A., and M. Weretka. 2007. *State prices and arbitrage in thin financial markets*. Madison: University of Wisconsin-Madison.

Chan, L., and J. Lakonishok. 1993. Institutional traders and intraday stock price behavior. *Journal of Financial Economics* 33: 173–199.

Chan, L., and J. Lakonishok. 1995. The behavior of stock price around institutional trades. *Journal of Finance* 50: 1147–1174.

DeMarzo, P.M., and B. Urošević. 2006. Ownership dynamics and asset pricing with a large shareholder. *Journal of Political Economy* 4: 145–174.

Dubil, R. 2002. Optimal liquidation of venture capital stakes. *Journal of Entrepreneurial Finance and Business Ventures* 7(2): 56–81.

Easley, D., and M. O'Hara. 1987. Price, trade size, and information in securities markets. *Journal of Financial Economics* 19: 69–90.

Engle, R., and R. Ferstenberg. 2007. Execution risk. *Journal of Portfolio Management* 33: 34–45.

Foster, F.D., and S. Viswanathan. 1996. Strategic trading when agents forecast the forecasts of others. *Journal of Finance* 51(4): 1437–1478.

Glosten, L., and P. Milgrom. 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 13: 71–100.

Grossman, S.J., and M.H. Miller. 1988. Liquidity and market structure. *Journal of Finance* 43(3): 617–637.

Ho, T., and H.R. Stoll. 1981. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics* 9: 47–73.

Holden, C.W., and A. Subrahmanyam. 1996. Risk aversion, liquidity, and endogenous short horizons. *Review of Financial Studies* 9(2): 691–722.

Huberman, G., and W. Stanzl. 2004. Price manipulation and quasi-arbitrage. *Econometrica* 74(4): 1247–1276.

Keim, D., and A. Madhavan. 1995. The anatomy of the trading process: Empirical evidence on the behavior of institutional traders. *Journal of Financial Economics* 37: 371–398.

Keim, D., and A. Madhavan. 1996. The upstairs markets for large-block transactions: Analyses and measurement of price effects. *Review of Financial Studies* 9: 1–39.

Keim, D., and A. Madhavan. 1998. Execution costs and investment performance: An empirical analysis of institutional equity trades, Rodney L. White Center for Financial Research Working Paper, 09-95.

Kyle, A.S. 1985. Continuous auctions and insider trading. *Econometrica* 53: 1315–1336.

Kyle, A.S. 1989. Informed speculation and imperfect competition. *Review of Economic Studies* 56: 517–556.

Mitchell, M., L.H. Pedersen, and T. Pulvino. 2007. Slow moving capital. *American Economic Review, P&P* 97(2): 215–220.

Perold, A. 1988. The implementation shortfall: Paper versus reality. *Journal of Portfolio Management* 14: 4–9.

Pritsker, M. 2005. Large investors: Implications for equilibrium asset, returns, shock absorption, and liquidity, Finance and Economic Discussion Series 2005-36, Board of Governors of the Federal Reserve System.

Rostek, M., and M. Weretka. 2007. *Frequent trading and price impact in thin markets*. Madison: University of Wisconsin-Madison.

Rostek, M., and M. Weretka. 2008. *Small double auctions*. Madison: University of Wisconsin-Madison.

Subramanian, A., and R. Jarrow. 2001. The liquidity discount. *Mathematical Finance* 11: 447–474.

Urosevic, B. 2005. Moral hazard and dynamics of insider ownership stakes. Universitat Pompeu Fabra, Economics Working Papers, 787.

Vayanos, D. 1999. Strategic trading and welfare in a dynamic market. *Review of Economic Studies* 66: 219–254.

Vayanos, D. 2001. Strategic trading in a dynamic noisy market. *Journal of Finance* 56(1): 131–171.
Weretka, M. 2006. *Endogenous market power*. Madison: University of Wisconsin-Madison.

Third World Debt

François Bourguignon

Abstract

Excessive dependence on foreign savings is a threat. But governments can lead by example by securing fiscal surpluses. As well, appropriate financial sector, tax and other microeconomic policies can help stimulate private domestic savings. Under these conditions, foreign borrowing can be a healthy complement to domestic savings.

Keywords

Capital account liberalization; Consumption smoothing; Contract enforcement; Debt crises; Debt intolerance; Debt overhang; Debt relief; Debt relief Laffer curve; Dollarization; Exchange rate policy; Fixed exchange rates; Floating exchange rates; Heavily indebted poor countries; International reserves; Liquidity constraint; Solvency constraint; Sovereign debt; Third World debt

JEL classification

O1

Many developing countries have encountered difficulties in managing their foreign debts in the nineteenth and twentieth centuries. A number of crisis episodes are reviewed by Fishlow (1985). Some of them, such as the Baring crisis of 1890, threatened the stability of the international financial system. Others were confined more specifically to a single country with a dispersion of creditors on bonded debt, as was the recent case of the Argentine default of 2002. While there is a long history of 'Third World' debt accumulation and subsequent defaults, the frequency of debt crises in developing countries has increased dramatically since the 1982 Mexican debt crisis. The focus of this article is the episodes since the mid-1980s, and the economic literature that emerged to analyse these events.

Much of the attention of the international community on Third World debt during the 1980s and early 1990s was focused on middle-income countries. During the mid- to late 1990s, debt relief for highly indebted poor countries (HIPCs) increasingly occupied the attention of policymakers around the world, as debt relief became a cause célèbre for a number of international NGOs. An increasing share of overseas development assistance in the new millennium has been devoted to cancelling the official credits of bilateral and multilateral donors through the Highly Indebted Poor Country 'HIPC' Initiative.

Readers who want more detailed accounts should consult the volumes edited by Smith and Cuddington (1985), Sachs (1989) and Husain and Diwan (1989), as well as the books by Cohen (1991) and Cline (1995). On debt relief for low-income countries, a detailed discussion is provided by Birdsall and Williamson (2002), or one can consult the volume edited by Addison et al. (2004). On recent analyses on the origins of debt crises as well as links to exchange rate policy, the reader is referred to Calvo and Reinhart (2002), Reinhart et al. (2003) and Eichengreen et al. (2003). Cline (1995, ch. 4) provides a critical review of the theory of sovereign borrowing.

General Analytic Framework

The neoclassical view of sovereign borrowing by developing countries is straightforward. Developing countries have lower capital stocks and resulting higher returns to capital than the high-income countries. These circumstances provide both borrowers and lenders with the incentive to engage in mutually beneficial debt transactions. The experience with sovereign defaults over the last two centuries leads to other considerations. Debt contracts across international borders do not

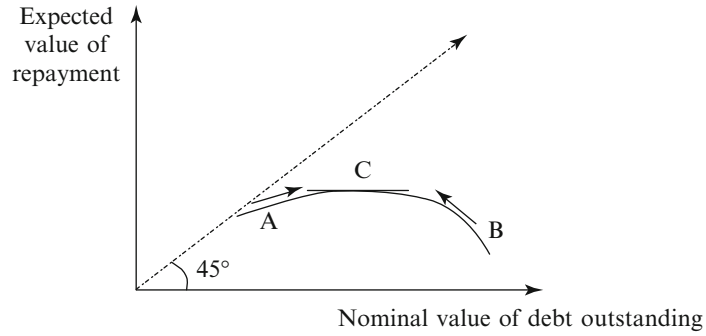
fall under a clear legal jurisdiction for enforcement, so the incentives for repayment on the part of the borrower must be understood.

Eaton and Gersovitz (1981) developed a model where developing countries borrow for consumption-smoothing purposes. Under certain circumstances, the desire for continued access to foreign credit for these purposes can be an adequate incentive for borrowers to pay, an incentive that is balanced against the one-off gains from defaulting. Creditors, understanding these incentives, will lend only up to the point at which the marginal costs and benefits of lending are equal. Clearly, this assumes that there is full information about these costs and benefits and coordinated action on the part of the creditors.

Cooper and Sachs (1985) provide another seminal analysis on foreign borrowing where the repayment question is divided into three components, or risks: illiquidity, insolvency and repudiation. The analysis looks at the optimal borrowing path for a debtor nation, taking into account constraints associated with the aforementioned three risks. For example, external solvency is limited by the present discounted value of future trade surpluses. If these surpluses happen to grow at a slower pace than the rate of interest, then the country faces a solvency constraint. The liquidity constraint may arise when, in a given year, the external trade surplus is insufficient to cover debt service, perhaps due to a short-run change in the terms of trade or other external factors. (Later analyses of solvency and liquidity would analyse explicitly an additional constraint: the internal transfer of resources from the domestic private sector to the government debtor. This issue is particularly troublesome in the context of devaluations to improve the trade balance – the external transfer constraint – which lead to a higher domestic currency tax bill that must be raised to effect the internal transfer from agents to the state; see Dornbusch 1988.) Despite long-run solvency, regulatory issues (for example, limited exposure to a single borrower) or coordination problems among creditors may lead individual banks to refuse to extend loans to cover that debt service. Repudiation risk, similar to the Eaton and Gersovitz concept discussed above, is another reason for creditors to limit loans, even when a negotiated solution to repudiation is anticipated.

Indeed, another critical question from the theoretical literature is the point at which debt relief, or debt forgiveness, becomes mutually beneficial to debtors and creditors. Krugman (1989) provides a simple definition of debt overhang: the expected present value of the resource transfer available to service debt is less than the value of debt. This sounds generally like an insolvency condition, expressed in expected present value terms, and the 'overhang' is the degree of excess debt that cannot be repaid under current expected conditions. Incentive problems then arise for overcoming the debt overhang. First of all, the existence of the overhang implies that there is a high marginal tax rate on domestic investment, as much of the return to investment will have to be captured by the sovereign to close the debt overhang. This poses an incentive problem for the sovereign attempting to adjust taxes and exchange rates to increase the capacity to pay: much of the gain from adjustment will accrue to the creditors. Thus, from the debtor's point of view, there are strong incentives to lobby for a reduction in principal, or a reduction in interest rates, as the only effective means to restore growth and overcome the crisis.

Understanding the dilemma faced by the debtor, creditors may also benefit from debt reduction. At high enough levels of debt, the probability of outright default increases rapidly enough that the expected total repayment to creditors may begin to fall as the overall face value of debt increases. In other words, there is a debt relief Laffer curve (Krugman 1989), as described in Fig. 1. At very low levels of indebtedness, the sovereign is expected to pay in full, and there is no discount on secondary markets. This is represented by the 45° line in Fig. 1: expected repayment equals the face value of the debt outstanding. As indebtedness levels increase, the probability of eventually facing a repayment problem increases. At point A, expected repayment is now below the face value of debt, and a secondary market discount would appear. However, at this stage it makes sense for creditors to continue lending – or even with 'defensive'



Third World Debt, Fig. 1 The debt relief Laffer curve

lending in the face of an external shock – in that an increase in the stock of debt outstanding still leads to an increase in expected repayment. This is no longer the case beyond point C. At point B indebtedness levels have reached the point where additional lending reduces expected repayment. Debt forgiveness, reducing the face value of debt outstanding, would then actually lead to an increase in the expected value of repayment. Yet this raises an issue of coordination among creditors since the value to the collective creditor community of reducing the debt by one dollar is higher than the average price (expected value) which represents the value of that operation to the individual investor (Bulow and Rogoff 1988). As a result some agreement must be reached among creditors on how to forgive debt, including 'burden-sharing' arrangements.

In summary, when debt becomes large enough, the prospects for simply growing out of the problem are dim. 'Debt overhang' can tax investment and limit debtors' incentives to engage in policy reforms to restore solvency, or even limit their incentive to continue to honour their obligations. As secondary market prices fall, the price for purchasing a claim on a debtor deviates greatly from the price a debtor has to pay to retire a claim via regularly scheduled debt service. An alternative is to use international reserves, if available, to purchase debt on the secondary market, thus lowering the cost of retiring the claim. Bulow and Rogoff (1988) point out, however, that the price paid on the secondary market is still too high, in that it represents the average price of debt, and ignores the marginal increase in expected repayments.

This analytical framework helps one understand the key issues facing both policy makers and creditors who are trying to realize mutually beneficial exits from a debt crisis. One overriding complication that emerges in a real crisis is uncertainty about the values of the parameters that define whether a country is facing a liquidity or a solvency problem, or if it is simply unwilling to pay. Krugman (1989) and others have asserted that, in an uncertain world, the distinction between liquidity and solvency is 'misleading': liquidity problems arise precisely because there is some reasonable probability that the country is insolvent. The brief recounting below of the history of recent debt crises reveals the difficulty in determining these values in the midst of a crisis. This uncertainty leads to divergent views on the best solution to a particular crisis episode.

This analytical framework continues to be useful. As noted below, the debt crises of the 1990s inspired a broader literature that examines how capital flow reversals, the currency structure of bank or corporate balance sheets and exchange rate policies all interact in causing the debt crises of this later period.

Evolution of External Debt: 1980s to the Present

Table 1 describes the evolution of the external debt of developing countries over the 1980–2003 period, in constant 2003 dollars. Two noteworthy trends are important for the discussion of the evolution of debt crisis episodes.

Third World Debt, Table 1 Total developing country external debt, 1980–2003

	1980	1985	1990	1995	2000	2003
*Total external debt**	1093.1	1421.4	1747.1	2379.7	2436.6	2553.0
Short term	269.6	266.6	310.8	490.0	401.2	508.6
Medium and long term	823.4	1154.8	1436.3	1889.8	2035.4	2044.4
Owed by public sector	719.8	1085.0	1405.4	1698.6	1517.4	1556.2
Owed by private sector	373.3	336.4	341.7	681.3	919.4	997.1
Owed to public creditors	342.9	510.1	803.4	1046.8	896.4	933.0
Owed to private creditors	750.2	911.3	943.8	1333.1	1540.4	1620.3
Memo items						
Share owed to private creditors	0.69	0.64	0.54	0.56	0.63	0.63
Share owed by private sector	0.34	0.24	0.20	0.29	0.38	0.39
Total external debt/GDP	0.18	0.28	0.32	0.38	0.38	0.37
Total external debt/exports (goods and services)	0.90	1.56	1.53	1.55	1.29	1.11

*Constant 2003 dollars, billions
Source: World Bank (2005a); author's calculations

First of all, we see a shift towards debt owed to 'public creditors' – that is, official bilateral and multilateral institutions – during the mid- to late 1980s, as the Baker and Brady Plans responded to the debt crises of the early 1980s. Over the same period, the share owed by the private sector in developing countries declined, as developing country governments assumed the liabilities of the private sector following the crisis. Later, during the early 1990s, we see an acceleration in the growth of the total debt stock, led by a greater role of both lending by private sector creditors and borrowing by private sector borrowers. This is not surprising, given that many developing countries pursued capital account liberalization at the start of the 1990s.

Debt Crises of the 1980s

The five paragraphs immediately following have been taken with some modification from Kenen (1992).

Commercial banks and other private institutions did not lend extensively to developing countries in the 1950s and 1960s. The build-up of debt that led to the crisis of the 1980s began in 1974, after the increase in the price of oil which followed the Arab–Israel war in October 1973. Some of the oil-exporting countries ran huge current account surpluses and deposited the proceeds with foreign commercial banks, and many oil-importing countries ran huge deficits. The imbalance was heavily concentrated on the developing countries because the industrial countries moved into recession in 1974–1975, reducing their total imports and focusing the global 'oil deficit' on the oil-importing developing countries. The economic environment of the 1970s was very conducive to large borrowing. Nominal interest rates were low, and real interest rates were negative. Many developing countries were posting higher growth rates of output and exports than in earlier decades, which held down the ratios of debt to output and of debt-service payments to exports. As noted in the analytical framework above, under these conditions debt accumulation is bound to appear sustainable.

Furthermore, the banks had protected themselves from two of the three risks facing them. First, the deposits of the oil-exporting countries were denominated mainly in US dollars, the currency in which they were paid for their oil, and most of the banks' own loans to developing countries were likewise dollar-dominated, so the banks were protected from exchange rate risks. Second, the interest rates charged on most of the loans were based on the London Inter-Bank Offered Rate (LIBOR), the interest rate paid on most of the deposits, and were adjusted frequently to maintain the spread over LIBOR, so the banks were protected from interest rate risks.

Yet the way that banks protected themselves from interest rate risk increased their exposure to default risk by linking the debtors' obligations to very volatile interest rates. Furthermore, the link to LIBOR, combined with the debtors' dependence on export earnings from products with cyclically sensitive prices, made it very likely that the debtor countries would prosper or suffer together, reducing the protection usually afforded by diversification. Finally, the risk of each bank's exposure to a particular country depended on the country's total debt, which depended in turn on the volume of loans made by other banks and on the amount of debt owned by all of the country's borrowers. In brief, there were externalities all over the place, but no one was paying attention to them.

The country which set off the crisis was just becoming an oil exporter rather than an oil importer. Mexico had borrowed heavily in the expectation of big oil exports, and much of its external debt bore floating interest rates. When interest rates rose sharply in the early 1980s, after the tightening of monetary policy in the United States, Mexico's debt-service payments soared. Together with capital flight, produced by expectations of devaluation, the increase in interest payments depleted Mexico's foreign-exchange reserves, and Mexico had to suspend those payments in August 1982.

The crisis was caused by a combination of the externalities on the creditor side, discussed above, and lax fiscal policies on the debtor side. In addition, countries with more closed trade regimes and overvalued exchange rates confronted the greatest difficulties in adjusting to the changes in the external environment.

By 1985, 15 countries were identified as highly indebted and requiring coordinated assistance from the international community, and these 15 were later extended to 17. In terms of magnitudes, the average debt–export ratio of this group peaked at 384% in 1986 (Cline 1995).

On the side of the creditors, there were serious concerns about a potential systemic crisis. Exposure rates reached remarkably high levels during the lending boom to developing countries. Sachs (1989) reports that at the end of 1982 the exposures of nine major banks to developing countries had reached a staggering 289% of bank capital, with much of this exposure concentrated in several Latin American countries. Prudential regulations on the concentration of exposure to individual borrowers did not avert this problem, since banks listed individual public agencies or state enterprises within each country as separate borrowers.

Policy Response: The Baker and Brady Plans

The initial response to the debt distress of the early 1980s was to coordinate efforts for new lending. Creditors largely comprised a group of large international banks. Payment arrears were viewed as primarily a problem of illiquidity. International economic conditions had turned against a number of heavily indebted emerging markets: sharply rising interest rates, declining prices for a number of commodities amidst a global economic recession. Coordinated efforts to extend new financing would allow these countries to honour their obligations until the external situation improved and structural reforms implemented. Projection models (for example, Cline 1985) suggested that, with feasible policy reforms and improving external conditions, most debtors could make good use of the new money and emerge as strong solvent sovereigns within a reasonable time period. The IMF would coordinate new lending.

These coordinated efforts were combined with targets for banks' and international financial institutions contributions under the 1985 'Baker Plan', named after James Baker, the US Treasury Secretary of that time. Private banks would provide $20 billion in loans, to be matched by a similar (gross) amount by the multilateral banks (Cline 1995).

Under the Baker Plan, substantial lending occurred, but less than the original commitments, partially due to the absence of IMF agreements in some defaulting countries. Some of the new lending was used for buybacks of existing debt, policy conditions were not always met on the recipient side and, in many countries, arrears on new loans began to accumulate. Meanwhile, by the late

1980s many creditor banks had been able to accumulate provisions against their losses. As a result, debt reduction or forgiveness would not cause the same degree of stress on the financial system as had been feared during the early years of the crisis. By 1988, a number of banks were used a 'menu approach', accepting concessional exit bonds or contributing to the new lending strategy. Also, by the late 1980s, secondary market prices had declined enough that banks that chose to exit via secondary markets were accepting large losses. Any coordinated reduction plan that would offer a reduction less than the secondary market discount would be viewed as favourable. A number of countries did emerge from the period of the 'new money' strategy without resorting to debt reduction: South Korea, Indonesia, and Turkey, are several prominent cases. Chile was a notable example from the Latin American region.

Under the leadership of Nicholas Brady, the new US Secretary of the Treasury, the 'Brady Plan' (1989–1994) was initiated with a focus on 'voluntary' debt reduction; however, still suggesting a menu of options including rescheduling for those banks/country cases where it might be of interest. The target for debt reduction was $70 billion. US government bonds were offered as collateral for new bonds used for debt reduction purposes, thus further attracting the creditors. Eventually, about $60 billion of debt was forgiven by 1994. In general, debt reduction averaged about 30% for private bank debts restructured; however, since official debt was about half the debt outstanding, total debt reduction amounted to about 15% of outstanding debt (Cline 1995). Renewed economic growth in the early 1990s, accompanied by renewed market access, seemed to indicate that the programme was a resounding success.

Debt Crises of the 1990s

By the end of the Brady Plan period, however, a new round of debt distress had appeared appear in emerging markets. During the late 1980s and early 1990s, many developing countries removed restrictions to capital inflows and liberalized their domestic financial sectors. As noted above, borrowing by the private sector led to the accumulation of debt, as did borrowing from private sector creditors. Another important change that emerged during the 1990s was the shift to bond issues, as opposed to commercial bank loans. This change obviously led to a greater dispersion and variety of creditors.

The new cycle of debt distress and crises again started with Mexico. In March 1994 an initial run on reserves occurred shortly following the assassination of the leading political candidate and a reversal of capital inflows which had been quite strong for a number of years. Mexico had been running large current account deficits during these years, in the context of a crawling band exchange rate policy, and international interest rates were also rising. In the aftermath of this first run on reserves, the central bank expanded domestic credit and it offered long-term bondholders new short-term, dollar-denominated *Tesobonos*. There was a fear of allowing interest rates to rise due to weaknesses in the banking system (Lustig 2001). By December, during the last month of the Salinas administration, the government was encountering increasing difficulty in rolling over *Tesobonos*. An attempted 'controlled devaluation' of 15% was followed by a renewed run on reserves. The monetary authorities were left with no option but to allow the currency to float.

Some authors (for example, Dornbusch and Werner 1994) attribute the Mexican crisis to a traditional problem of an overvalued fixed exchange rate, with persistent current account imbalances and the corresponding accumulation of debt. With the increasing indebtedness, shorter terms and higher interest rates were demanded by the market, further increasing the likelihood of default. Devaluation would also increase indebtedness ratios relative to domestic resources. Others (Calvo and Mendoza 1996) place greater emphasis on stock imbalances – namely, money balances, short-term debt and gross reserves – and sharp adjustment driven by global capital markets with 'herding' behaviour on the part of foreign bondholders. From this point of view, Mexico suffered 'cruel punishment' for the 'petty crime' of attempting a controlled devaluation to correct a

modest balance of payments disequilibrium. This debate again reflected some of the difficulties in determining whether the country was insolvent or merely facing a liquidity problem that could be resolved by some coordination among creditors, as discussed in the theoretical framework above.

Indeed, the 'punishment' was heavy. GDP declined by more than 6% in 1995, real wages declined sharply in the aftermath of the devaluation, and poverty rates increased by about seven or eight percentage points. Half a decade would pass before poverty rates returned to pre-crisis levels.

In early 1995, an international support programme to finance the gap in Mexico's balance of payments was arranged of the order of some $50 billion in commitments from the IMF and bilateral sources – mainly the United States. The Fund programme of about $18 billion was the largest lending programme in IMF history at that time. The floating of the currency, along with improved trade possibilities due to the implementation of NAFTA, and these official financing flows allowed Mexico to return to growth averaging about 5.5% over the 1996–98 period.

Possibly through contagion from the Mexico crisis, but also for domestic reasons, Argentina suffered a sharp withdrawal of capital and loss of international reserves in early 1995. There, too, an international support package was arranged, and robust growth was restored in 1996.

Mexico's crisis was then followed by several other financial crises in the late 1990s and into the new millennium, all involving either outright default or near default on substantial external liabilities. In East Asia, even South Korea – considered to be virtually a 'developed' country – was unable to avoid a substantial crisis based on the accumulation of short-term external liabilities of the private corporate and banking sector rather than government debt. Starting in Thailand, in the summer of 1997 the Asian financial crisis quickly spread by the end of the year to affect heavily Indonesia, Malaysia, Korea and the Philippines. Japan, Taiwan and Singapore also suffered through substantial problems in their financial sectors.

Each country affected by the East Asian crisis had particular circumstances leading up to the crisis; however, the most heavily hit countries had a number of common problems. Fixed exchange rates and either implicit or explicit guarantees by the government motivated private companies and banks to inadequately account for the risk involved in borrowing heavily in foreign currency. (The common features mentioned in this section are drawn from Kawai et al. 2001. Goldstein 1998, also emphasizes a combination of factors, including external imbalance and financial sector fragility due to mismatches.) There were widening current account deficits financed by short-term capital flows. Relatively unregulated financial systems conducted a credit boom with investment in low-return activities, along with substantial currency and maturity mismatches in their balance sheets. Corporations borrowed heavily, leading to exposure to interest and exchange rate shocks. Political uncertainty led to doubts about the continuation of existing exchange rate and other policies.

As in the case of Mexico, economists differ in assigning a direct cause of the crisis. There are those that emphasized fundamentals – along the lines of the common characteristics across countries mentioned above. Others (for example, Radelet and Sachs 1998) emphasize that the problem was one of illiquidity rather than insolvency, and foreign investor panic led to a withdrawal of the willingness to provide liquidity. Kaminsky and Schmukler (1999) provide empirical evidence to support the view that investors' reactions to news events – as measured by changes in the stock market – can be characterized as 'herding behaviour'.

The impact of the crisis was dramatic. The Thai and Indonesian economies shrank by 10.5 and 13.1% respectively, in real terms. The Korean and Malaysian economies shrank by 6.9 and 7.4%, respectively.

In response to the crisis, the IMF was working on numerous fronts to provide financial support to cover debt-servicing needs. Programmes were established in the main countries affected, and the World Bank also provided substantial fast-disbursing funds. Korea returned to borrower status with the World Bank, securing the largest individual loans in World Bank history, despite previous 'graduation' from borrower status.

Overall, including both multilateral and bilateral donors, $57 billion was committed for Korea, $35 billion for Indonesia and $17 billion for Thailand over the 1997–1998 period (Radelet and Sachs 1998).

In 1998, Russia was the next emerging market economy to encounter difficulties in servicing its external debt (Kharas et al. 2001). In August, 1998, the authorities announced a 90-day moratorium on external debt, as well as their intention to restructure all obligations due through the end of 1999. Within 3 weeks, the central bank floated the ruble with ruble–dollar rate subsequently tripling. The Russian authorities had attempted to sustain the exchange rate to secure inflation and to avoid the economic collapses of Asian countries that had devalued in the previous year. One month prior to the collapse, the authorities had arranged for a $22.6 billion external financing package led by the IMF. The IMF programme included measures to restore long-term fiscal balance. The short-term financing was also expected to restore foreign investor confidence and assist in a voluntary swap of short-term government ('GKO') bonds for longer term Eurobonds.

The impact of the crisis on Russia was strong, but fairly short-lived. GDP declined by 5.3% in 1998, but fully recovered in 1999. Headcount poverty (using the two dollar a day definition) rose from 23% in 1996 to 36% in 1998. By 2000 this measure of poverty had fallen to 24% again (World Bank 2005b).

It was feared at that time that the Russian crisis would contaminate other emerging countries. In effect, the lack of confidence of investors resulted in capital outflows in several countries. Combined with domestic events, the situation became somewhat serious in Brazil early in 1999. But the adoption of a floating exchange rate as well as tightening fiscal measures re-established confidence and reversed capital outflows.

Argentina, too, was contaminated by the Russian default. Its sovereign debt experienced a sharp rise in spreads, rising to nearly 15% over equivalent US bonds in September 1998. By the end of 1998, bond spreads had declined to below 10%; however, the country had slipped into a period of stagnation and eventual recession that lasted through mid-2001. A combination of lingering doubts among investors about the hard currency peg (a quasi-currency board arrangement), national elections in 1999, and falling international commodity prices contributed to the malaise. By mid-2001 confidence had reached a low point, Argentines were pulling (mostly dollarized) deposits out of banks and the central bank was losing reserves at a rapid pace. In the middle of the year, a 'mega swap' was conducted to prolong maturities without any reduction in the stock outstanding. This 'market-based' approach failed to stem the loss of reserves. New resources of approximately $6 billion by the IMF in August also failed to make a difference. In December, the government declared a freeze on deposits in a last-ditch effort to stem the haemorrhaging. By the end of the month, facing escalating street protests, the president would resign. In early January 2002 an interim president declared unilateral suspension of payments on foreign public debt. The exchange rate would soar from the fixed 1:1 with the dollar (which had dated back to 1991) to over three ARS per dollar.

This latest crisis was particularly dramatic. The population had treated the 1:1 fixed exchange rate with the dollar as a national institution. The vast majority of deposits, loans and business contracts were set in dollars. A combination of terms of trade shocks and declining investors' confidence in politicians' willingness or ability to sustain the system would be accompanied by a painful, gradual, deflationary adjustment of the real exchange rate leading to the eventual collapse in early 2002.

Resolution of the debt default was also particularly problematic. There were many series of bond issues in various international legal jurisdictions and numerous issues that had been passed on by underwriters to the retail level. Relations between the government and multilateral lenders also reached a stand-off, with little or no 'fresh money' provided, and brief periods of suspension of payments. For private creditors, the government took a hard-line position: an offer of approximately 25 cents on the dollar, in present value terms – a level lower than the value implied by secondary markets at the time. Some 100 billion dollars were eligible to be exchanged for a menu

of bonds, both par and discount and also in either domestic or foreign currency. By the time of the actual transaction international interest rates had declined, so that the final offer presented a present value of just over 30 cents to the dollar of face value, a level close to prevailing secondary market discounts. This transaction represented the sharpest discounts on sovereign debt exchange in history. Just over three-fourths of the defaulted bonds were eventually exchanged, leaving a fairly substantial community of 'holdouts'.

GDP declined by an average of 2.9% per year over the 1999–2001 period. In 2002 the economy shrank by nearly 11%. Headcount poverty rates (national poverty line) more than doubled to around 50% in 2002. Thanks to a robust recovery, GDP in real local currency terms would finally return to 1998 levels in mid-2005. The poverty rate would decline to around just under 40%.

Debt and Economic Policies: Analysis of the Crises of the 1990s

As noted in the analytical framework above, indebtedness strategies cannot be viewed in isolation from other economic policies. The ability to pay is determined by the growth of exports and the effectiveness of the debtors' tax system to secure resources for debt repayment. Clearly export growth is affected by both microeconomic and macroeconomic policies, and in particular, the exchange rate regime. Vulnerabilities in the financial system, due to lax regulation and supervision, can trigger bank withdrawals, capital flight, and exchange rate movements that dramatically shift a country from 'apparent' solvency to immediate insolvency.

A new nomenclature has developed in the recent literature. Many developing countries commit the 'original sin' (Eichengreen et al. 2003) of borrowing in foreign currency. Dollar-denominated liabilities, or even broader dollarization of the domestic financial systems, result in a 'fear to float' (Calvo and Reinhart 2002) the nominal exchange rate. Countries with fixed exchange rate regimes supported by large capital inflows face the risk of 'sudden stop' (Calvo et al. 2003) of these inflows, leading to distressed debt restructuring or default. The prevalence of debt crises in emerging economies with debt to GDP ratios below OECD averages has led some analysts to diagnose these countries as suffering from 'debt intolerance' (Reinhart et al. 2003).

HIPCs

In parallel to these events involving countries borrowing from private markets, there was growing concern that the poor economic performance of many low-income countries was due to excessive indebtedness. A fundamental difference for these countries was that they had virtually no access to private markets for lending: debt stocks had been accumulated with official bilateral and multilateral lenders. There is no rollover risk and the behaviour of private lenders is not an issue. On the other hand, a number of economists in the 1980s began to emphasize debt overhang as a deterrent to growth in heavily indebted poor countries, or 'HIPCs', as they would come to be known. Others, supported by prominent international NGOs, emphasized that debt service detracted from the resources available for poor countries' governments to provide basic services for their people.

Debt relief for poor countries dates back to the establishment of the 'Paris Club' in the 1950s – the coordinating institution for bilateral creditors to agree on debt relief. Since the early 1980s, however, multilateral efforts to provide debt relief have intensified. Easterly (1999) provides a review of debt relief for poor countries over the previous two decades. The 1977–1979 meetings of the United Nations Conference on Trade and Development led to debt write-offs of some $6 billion for 45 lowincome countries. In the early and mid-1980s, World Bank reports on Africa increasingly emphasized the debt servicing difficulties of low-income countries. By the late 1980s, debt relief for these countries had entered the agenda of the annual G7 summits. The summit of 1988 in Toronto established a menu of debt relief terms to be offered to these countries. The multilaterals established special

programmes as well: the World Bank's Special Program of Assistance (SPA) to low-income Africa and the IMF's Enhanced Structural Adjustment Facility (ESAF). These facilities would provide fast-disbursing funds to assist in debt repayments.

During the 1990s low-income country debt relief gained further momentum. On the global political front, there was growing support for debt relief – including from religious leaders and high-profile figures in popular culture. The year 2000 marked a symbolic target date for debt relief and 'Jubilee 2000' political movements to promote debt relief spread across some 60 countries. According to Jubilee Research, a global petition for debt relief collected 24 million signatures. The 'Toronto terms' debt relief mentioned above was followed by a series of expansions of debt relief during international summits in the early 1990s. In 1996, the World Bank and the IMF established the Highly Indebted Poor Country (HIPC) Initiative, and in the same year, the Paris Club of bilateral creditors committed to 80% debt relief in net present value terms (Easterly 1999). In 1999, the HIPC initiative was 'enhanced' to provide faster and broader debt relief and to link debt relief with the elaboration of poverty reduction strategies (as countries were asked to prepare Poverty Reduction Strategy Papers, or PRSPs). These papers would be prepared in collaboration with, and with the support of, the World Bank and IMF. The IMF initiated a new lending facility called the Poverty Reduction and Growth Facility (PRGF) to provide interim financing to the HIPCs as they prepared these papers.

The technical criterion for a low-income country to be considered 'highly indebted' is either a net present value (NPV) of debt greater than 150% of exports or an NPV of debt greater than 250% of government revenues. HIPCs must pass through a process of several steps before receiving debt relief. A 'Decision Point' is reached when a country displays a minimum degree of macroeconomic stability, has cleared any arrears in debt service, and has prepared an Interim PRSP. The World Bank and IMF also prepare a debt sustainability analysis at this stage. At the decision point, an initial level of debt relief is granted 'conditionally' upon successful passage to the 'Completion Point' stage.

During the 'interim' period between the decision point and completion point, the IMF provides financing under a PRGF and the World Bank may provide policy-based credits through the International Development Association (IDA) as well. If a country has achieved satisfactory performance under the agreed measures of a PRGF and the PRSP has been implemented for at least 1 year, then the country may reach the Completion Point. At this point, debt relief is granted 'irrevocably'. Additional debt relief – 'topping off' – may also be provided by multilateral and bilateral creditors.

As of mid-March 2005, 27 countries had reached or surpassed the Decision Point, of which 15 had reached the Completion Point, representing a total of approximately $32 billion (in NPV terms at the year of the Decision Point) of debt relief committed by multilateral and bilateral creditors (IDA–IMF 2005). Another 11 countries were at the pre-Decision Point stage. In NPV terms of 2004, the total expected cost is $58 billion. For the 27 Decision Point (or beyond) countries, debt stocks were reduced by two-thirds (see World Bank 2007).

The HIPC Initiative has generated both supporters and critics. A number of NGOs have complained that the process is too slow and onerous. Easterly (1999) emphasizes that debt relief is likely to be followed by a re-accumulation of debt unless countries change their long-run savings preferences. There have been complaints that the targets are too 'ad hoc' and the 'criteria for debt relief too narrow' (Addison et al. 2004). Others have stated that the programme has generated false expectations in that the real resource transfer involved is limited because these debts would have been 'rolled over' indefinitely (Cohen 2001). On the other hand, there have been proposals to expand debt relief, with deeper relief for current HIPCs: broadening the scope to include more low-income countries, safeguarding countries from returning to unsustainable debt levels via a new contingency facility, and shifting the focus of future development assistance from loans to grants and with more multilateral coordination of these aid flows (Birdsall and Williamson 2002).

At the annual meetings of the IMF and World Bank in 2005, donor countries committed to additional debt relief, mostly via covering capital losses to the multilateral institutions if they cancel their own loans to HIPCs.

Looking Ahead: The 'New Financial Architecture'

Much of the focus of a new system for international debt flows is on how to make crisis resolution less costly for both debtors and creditors. The IMF proposed a new Sovereign Debt Restructuring Mechanism (SDRM) (Krueger 2002), that would attempt to capture some of the features of corporate debt restructuring in developed countries. The main features of this proposal were fourfold: (*a*) 'majority restructuring' – a mechanism whereby a qualified majority of bondholders could commit the minority to a restructuring agreement; (*b*) 'stay on creditor enforcement' – a period during which creditors would not be able to pursue litigation to receive payment while attempts are made to reach a majority restructuring agreement; (*c*) protection of creditor interests – guarantees that debtors will not make payments to 'non-priority' creditors and that appropriate policies are pursued by the debtor so as not to reduce asset values (an IMF programme might be the means of supporting the latter); and (*d*) priority financing – a mechanism to reach agreement on 'new money' to facilitate the debt restructuring process, with new money receiving 'senior status' relative to old debt.

In principle, the SDRM addresses directly many of the incentive and coordination problems discussed above. To date, little progress has been made in implementing the proposal, with perhaps the exception of the first clause. Collective action clauses allow for a pre-defined majority or supermajority of bondholders to commit to a debt restructured with a distressed debtor. Collective action clauses are already common in many debt contracts issued under British law; however, this spontaneous market response does not address the aggregation issue. The clauses are attached and specified for individual bond issues, and thus they would not assure that all privately held debt could come under a single rule. Some other form of international agreement or institution would be needed to solve the aggregation issue.

Lessons from the Crises: Domestic Debt and Fiscal Management

A combination of accumulating short-term external debt and heavily managed, or fixed, exchange rates can be fatal, leading to debt distress and subsequent economic strife. Many governments in developing countries are, as of the early twenty-first century, rebalancing their balance sheets towards debt denominated in domestic currency. By de-linking public debt stocks from the exchange rate, countries are gradually overcoming their 'fear of floating' the exchange rate. With more competitive exchange rates, a number of previous defaulters are also running balance of payments surpluses and improving their net foreign asset positions. All of these changes are occurring gradually, leaving them exposed to sudden shocks in the external environment.

Fiscal management remains an issue, however, in many of these countries, and it can limit growth prospects. The shift to borrowing locally crowds out domestic private borrowers, especially in those cases where domestic financial systems are still small following collapses during previous debt crisis episodes. In addition, a number of countries still have domestic financial systems with assets and liabilities denominated in foreign currency: another motivation for 'fear of floating'.

In the end, an excessive dependence on foreign savings is a threat, as taught to us by the fundamentals of the intertemporal external balance constraints discussed earlier. In fact, many analysts have turned their attention to global imbalances involving the persistent current account deficits of the Unites States as a potential source of future disturbances that could affect those developing countries that are gradually trying to unwind their net external liabilities (see, for example, Goldstein 2005). Clearly, governments can lead by example by securing fiscal surpluses (Gill and Pinto 2005, conclude from their review that the first major lesson of the debt crises is 'that paying

attention to the government's intertemporal budget constraint.. is vital'). As well, appropriate financial sector, tax and other microeconomic policies can help stimulate private domestic savings (Cohen 1992, concludes that a shift up in the capital stock needs to be accompanied by a level of saving consistent with this larger capital stock in order to sustain growth; he notes that the recipient country must also have an unusual combination: a relatively high endowment of human capital and relatively low physical capital). Under these conditions, foreign borrowing can be a healthy complement to domestic savings.

See Also

► Financial Structure and Economic Development

Bibliography

Addison, T., H. Hansen, and F. Tarp. 2004. *Debt relief for poor countries*. Basingstoke: Palgrave Macmillan.

Birdsall, N., and J. Williamson. 2002. *Delivering on debt relief: From IMF gold to a new aid architecture*. Washington, DC: Center for Global Development and the Institute for International Economics.

Bulow, J., and K. Rogoff. 1988. The buyback boondoggle. *Brookings Papers on Economic Activity* 1988(2): 675–98.

Calvo, G.A., and E.G. Mendoza. 1996. Petty crime and cruel punishment: Lessons from the Mexican debacle. *American Economic Review* 86: 170–5.

Calvo, G.A., and C. Reinhart. 2002. Fear of floating. *Quarterly Journal of Economics* 117: 379–408.

Calvo, G.A., Izquierdo, A. and Talvi, E. 2003. *Sudden stops, the real exchange rate, and fiscal sustainability: Argentina's lessons*. Working Paper No. 9828. Washington, DC: NBER.

Cline, W.R. 1985. International debt: From crisis to recovery. *American Economic Review* 75: 190–8.

Cline, W.R. 1995. *International debt reexamined*. Washington, DC: Institute for International Economics.

Cohen, D. 1991. *Private lending to sovereign states: A theoretical autopsy*. Cambridge, MA: MIT Press.

Cohen, D. 1992. The debt crisis: A post mortem. In *NBER macroeconomics annual*, vol. 7, ed. O. Blanchard and S. Fischer. Cambridge, MA: MIT Press.

Cohen, D. 2001. The HIPC initiative: True and false promises. *International Finance* 4: 363–80.

Cooper, R.N., and J.D. Sachs. 1985. Borrowing abroad: The debtor's perspective. In *International debt and the developing countries*, ed. G.W. Smith and J.T. Cuddington. Washington, DC: World Bank.

Dornbusch, R. 1988. Our LDC debts. In *The united states in the world economy*, ed. M. Feldstein. Chicago: University of Chicago Press.

Dornbusch, R., and A. Werner. 1994. Mexico: Stabilization, reform and no growth. *Brookings Papers on Economic Activity* 1994(1): 253–315.

Easterly, W. 1999. *How did highly indebted poor countries become highly indebted? Reviewing two decades of debt relief*. Policy Research Working Paper No. 2225, World Bank.

Eaton, J., and M. Gersovitz. 1981. Debt with potential repudiation: Theoretical and empirical analysis. *Review of Economic Studies* 48: 289–309.

Eichengreen, B., Hausmann, R. and Panizza, U. 2003. *Currency mismatches, debt intolerance and original sin: Why they are not the same and why it matters*. Working Paper No. 10036. Cambridge, MA: NBER.

Fishlow, A. 1985. Lessons from the past: Capital markets during the 19th century and the interwar period. *International Organization* 39: 383–439.

Gill, I. and Pinto, B. 2005. *Public debt in developing countries: Has the market-based model worked?* Policy Research Working Paper No. 3674, World Bank.

Goldstein, M. 1998. *The Asian financial crisis: Causes, crises and systemic implications*. Washington, DC: Institute for International Economics.

Goldstein, M. 2005. *What might the next emerging-market financial crisis look like?* Working Paper No. WP 05–7, Institute for International Economics.

Husain, I., and I. Diwan. 1989. *Dealing with the debt crisis*. Washington, DC: World Bank.

International Development Association and International Monetary Fund. 2005. *Heavily Indebted Poor Countries (HIPC) Initiative – Statistical update*. 4 April 2005. Online. Available at http://siteresources.worldbank.org/INTDEBTDEPT/ProgressReports/20446696/HIPCStatUpdate200504042.pdf. Accessed 11 Apr 2007.

Jubilee Research. Online. Available at http://www.jubileeresearch.org/about/about.htm. Accessed 11 Apr 2007.

Kaminsky, G.L. and Schmukler, S.L. 1999. *What triggers market jitters? A chronicle of the Asian crisis*. Policy Research Working Paper No. 2094, World Bank.

Kawai, M., Newfarmer, R. and Schmukler, S. 2001. *Crisis and contagion in east Asia: Nine lessons*. Policy Research Working Paper No. 2610, World Bank.

Kenen, P. 1992. Third World debt. In *The new palgrave dictionary of money and finance*, ed. P. Newman, M. Milgate, and J. Eatwell. London: Macmillan.

Kharas, H., B. Pinto, and S. Ulatov. 2001. An analysis of Russia's 1998 meltdown: Fundamentals and market signals. *Brookings Papers on Economic Activity* 2001(1): 1–67.

Krueger, A.O. 2002. *A new approach to sovereign debt restructuring*. Washington, DC: International Monetary Fund.

Krugman, P. 1989. Market-based debt-reduction schemes. In *Analytics of international debt*, ed. J. Frankel. Washington, DC: International Monetary Fund.
Lustig, N. 2001. Life is not easy: Mexico's quest for stability and growth. *Journal of Economic Perspectives* 15(1): 85–106.
Radelet, S., and J. Sachs. 1998. The East Asian financial crisis: Diagnosis, remedies, prospects. *Brookings Papers on Economic Activity* 1998(1): 1–74.
Reinhart, C., Rogoff, K. and Savastano, M. 2003. *Debt intolerance*. Working Paper No. 9908. Washington, DC: NBER.
Sachs, J. 1989. *Developing country debt and economic performance, volume I: The international financial system*. Chicago: NBER and University of Chicago Press.
Smith, G.W., and J.T. Cuddington. 1985. International borrowing and lending: What have we learned from theory and experience? In *International debt and the developing countries*, ed. G.W. Smith and J.T. Cuddington. Washington, DC: World Bank.
World Bank. 2005a. *Global development finance*. Washington, DC: World Bank.
World Bank. 2005b. *World development indicators 2005*. Washington, DC: World Bank.
World Bank. 2007. *Debt issues*. Online. Available at http://www.worldbank.org/debt. Accessed 11 April 2007.

Thompson, Thomas Perronet (1783–1869)

Murray Milgate and Alastair Levy

Keywords

Corn Laws; Currency; Ricardo, David P; Theory of rent; Thompson, Thomas P; Westminster Review

Appointed as the first Crown Governor of the British territory of Sierra Leone in 1808, Thompson was recalled under suspicion of financial impropriety in 1809. The real explanation for his departure, however, had more to do with the fact that the Sierra Leone Company (which had governed since 1790) found excessively disturbing Thompson's determination to rid the colony of an apprenticeship system whose features, as he saw it, were hardly different from those of slavery. The abuses which Thompson observed had developed, it should be noted, despite the fact that the Sierra Leone Company had been set up by anti-slavery philanthropists, including William Wilberforce and the economist Henry Thornton, with the intention of returning liberated slaves from the Americas to Africa (and, it was hoped, to illustrate the profitability of an African colonial trade not based on slavery).

On his return voyage to England, Thompson was the victim of an act of piracy. His vessel was boarded by the crew of a French corvette, and while its captain entertained Thompson, the British vessel was liberated of its cargo and provisions. Once safely back in England, Thompson applied to the Prime Minister (Lord Liverpool) for another official posting, but 'in case no other situation should present itself', he considered the possibility of single-handedly introducing the study of political economy into the University of Cambridge 'in order to provide a living for myself' (letter to E.P. Sells, January 1811, cited in Johnson 1957, p. 70). This idea was not entirely fanciful, since Thompson was a graduate of Queen's College (BA, seventh wrangler, 1802) and a fellow of that college. It transpired, however, that the first regular lectures on the subject at Cambridge were given by George Pryme in 1816, and Thompson instead reactivated his commission in the army (into which service he had switched from the navy in 1806). His command of a defeat at Muscat led to his being court martialled but acquitted (though with a reprimand for 'rashly undertaking the expedition with so small a detachment') in 1820.

In 1822 Thompson played an active role in the founding the *Westminster Review* (financed by a £4,000 advance from Bentham), which aimed to provide a radical alternative to the Tory *Quarterly Review* and the Whig *Edinburgh Review*. James Mill turned down the offer of its editorship, although he did contribute to its first number (in 1824) for which Thompson himself wrote an article on the 'Instrument of exchange'. This constituted his first published work on economics. Thompson was sole proprietor of the *Review* from 1829 until 1836 when, on his election to the reformed House of Commons for the seat of Hull (where he had been born on 15 March 1783), he transferred its ownership to William Molesworth.

In 1826 Thompson published the first of his longer tracts on economics, *An Exposition of Fallacies on Rent, Tithes, & etc*. Ostensibly an attack on the Ricardian theory of rent (indeed, Thompson retitled its second edition *The True Theory of Rent in Opposition to Mr Ricardo and Others*), John Stuart Mill described it as 'a striking exemplification of the mistakes of an ingenious, but not thoroughly informed mind' and claimed that, in fact, Thompson's 'theory of rent differs from that of Mr Ricardo only in the expression' (1828, pp. 178–9). However, Mill's claim is open to question. Thompson argued that rent was determined by 'the limited quantity of land in comparison with the competitors for its produce' (1826, pp. 8–9) – quite how this could be said to be essentially Ricardo's theory 'in different words' is difficult to see. Not only does it fail to distinguish between extensive and intensive rent, but it ignores altogether the effect on the conditions of production of wage goods of restrictions on the importation of corn which is the key to Ricardo's argument. The only observation that needs to be made about Mill's claim is, perhaps, that it may tell us rather more about his own contribution to the decline of Ricardian economics than about Ricardo's theory of rent. That Mill could advance such a claim within five years of Ricardo's death makes it less difficult to understand why many felt that 'little remained of Ricardo's theory' by the end of that decade.

In 1827 followed the *Catechism of the Corn Laws* which Mill pronounced 'one of the most useful works which have appeared in the present controversy' (1828, p. 186); an interesting judgement given that its opening section was based on the *Exposition*. The *Catechism* presents a list of 120 (later increased to 365) 'Proteus-like fallacies' and answers, and has been referred to as 'the arsenal whence the Anti-Corn Law League drew its best weapons' (Allibone 1871). There then followed, during the seven years he owned the *Westminster Review,* better than 100 articles for that periodical on subjects as diverse as the reform of the House of Lords and Catholic emancipation. Most of these were republished in his multi-volume *Exercises, Political and Other* in 1842. Also to be mentioned are his opinions on currency questions (for example, 1848), where he was an opponent of 'inflationist' proposals largely on the grounds of the redistribution against workers which he saw as part of the process. In the crisis of 1847, when the Bank Act of 1844 was again the subject of hot debate, Thompson wrote: 'I hold to my opinion that there will be mischief on the Currency question. I receive more half-mad pamphlets from Birmingham' (letter to J. Bowring, April 1847, quoted in Johnson 1957, p. 265). In 1852 he attempted to introduce into Parliament measures which would protect the value of the currency against depreciation due to new gold discoveries, but these were defeated.

There is much more that could be said of Thompson's remarkable career. He was a moral-force, class-alliance chartist (and was invited to participate in writing the draft act of parliament which was to become the People's Charter); he voted consistently with the Radicals when a member of parliament; he constructed, and published, a non-axiomatic system of geometry (Euclid without the axioms); and he invented an enharmonic organ which was exhibited at the Great Exhibition of 1851, where it received an honourable mention. He died at Blackheath on 6 September 1869.

Selected Works

1826. *An exposition of fallacies on rent, tithes, & etc*. London: Hatchard & Son. 2nd ed, retitled *The true theory of rent in opposition to Mr Ricardo and others*, 1826.
1827. *A Catechism on the corn laws; with a list of fallacies and answers*. London: J. Ridgway.
1842. *Exercises, political and other*, 6 vols. London: E. Wilson.
1848. *A Catechism on the currency.* London: E. Wilson.
1859. *Catechism on the Ballot; or a list of fallacies and the answers*. London: G. Brown.

References

Allibone, S.A. 1859–71. *A critical dictionary of English literature and British and American authors, living and*

deceased, 3 vols. Philadelphia: Childs & Peterson; London: N. Trübner & Co.
Johnson, L.G. 1957. *General T. Perronet Thompson: 1763–1869: His military, literary and political campaigns*. London: Allen & Unwin.
Mill, J.S. 1828. A review of the third edition of *Catechism on the corn laws. Westminster Review* 13: 169–187.

Thompson, William (1785–1833)

N. W. Thompson

Self-confessedly 'one of the idle classes', Thompson, the son of a merchant, was a substantial landowner with an estate in County Cork. He evinced an early interest in 'advanced' opinions which led to contacts with the St. Simonians in France and in 1822 his intellectual interests took him to London where he resided for a time with Bentham in Queen Square Place. Here he met some of the leading philosophical radicals and classical political economists of the day, such as Robert Torrens and James Mill. His sympathies were, however, enlisted by the expanding cooperative movement and by 1825 he was, in the words of J.S. Mill, 'the chief champion on the co-operative side' in a series of debates held in the metropolis on the subject of co-operation.

Thompson's major work of political economy, *An Inquiry into the Principles of the Distribution of Wealth* was published in 1824. In it, Thompson set out the principles which he believed should regulate economic life, namely the free direction of labour, voluntary exchanges and the use by labour of its entire product. These were the 'natural laws' which, if they prevailed 'would produce much happiness in any community' and it was the purpose of the *Inquiry* to examine the extent to which they did prevail first, under existing economic arrangements; secondly, in a truly competitive market economy; and thirdly, under a system of mutual cooperation.

For Thompson, existing economic arrangements were governed by 'absolute violence, fraud . . . the operation of unequal laws interfering with the freedom of labour . . . and the perfect freedom of voluntary exchanges'. This resulted in the appropriation of labour's product by 'a class of capitalists a class of rent or land-owners, and an always imperious class of idlers'. It was these classes who, through the coercive exercise of economic and political power, 'counteract (ed) the natural laws of distribution', 'forcing labour without a satisfactory equivalent'.

Yet in the *Inquiry* Thompson seems to have accepted that force and fraud were not a necessary feature of a competitive market economy. Rather they were aberrant intrusions which obstructed its equitable functioning. Render a market economy truly free and competitive and 'the natural laws of distribution' would prevail. Thus such an economy might play a distributive role which militated strongly against substantial material inequalities. As he wrote in the *Inquiry*,

> 'tis by means of the brutal expedients of insecurity . . . by the varied employments of force and terror . . . that the capitalist is enabled to keep down the remuneration of labour . . .. The mere competition of producers, if left to the natural laws of distribution . . . would be entirely of the exhilarating instead of the depressing species.'

It would act 'constantly to raise the remuneration of labour . . . while . . . at the same time cheapen the articles produced to society at large'. Untramelled competition would 'banish extremes of wealth and poverty' and society would enjoy 'blessings of equality comparable to those enjoyed under Mr. Owen's system of mutual co-operation by common labour.'

Yet for Thompson this was not sufficient and in the *Inquiry* he proceeds to press the case for communities of mutual co-operation where all would have an equal right to draw upon the products of co-operative labour and where the voluntary renunciation of personal rights to property would obviate any violation of the principle of security in the products of individual labour.

In the *Inquiry* communities were preferred largely because of their benign social and ethical consequences. Thus Thompson expressed anxiety over the moral tone of a society which 'retain (ed) the principle of selfishness . . . as the leading motive to action' and where there was, in

consequence, negligible scope for action motivated by benevolence and social concern. Such a society bred antagonism and conflict. The competitive pressures it unleashed threatened its social cohesion and harmony, while the moral corrosion it engendered was revealed in the transmutation of truth, sincerity benevolence and man himself into marketable commodities.

The publication of *Labor Rewarded* (1827) saw an even more emphatic rejection of the competitive market economy. Thus Thompson argued in this work that even in the non-slave states of America, were it existed in its purest actualized form, individual competition produced not only the moral and social evils of the kind detailed in the *Inquiry* but also general destitution and distress with labour, in many cases, securing a 'mere sufficiency of food and clothing'.

What seems to have provoked this more hostile attitude was Thompson's recognition that whatever the theoretical ideal, the reality was that 'free competition' had 'never yet practically meant anything else but a sham' for if it was to mean anything it required the existence of 'equal means of knowledge and skill, equal freedom of action, equal materials for production and accumulation, equal rights and duties' and equality of 'fortunes' when 'beginning the race of competition'. These preconditions had never been met and so the defence of competition became, in fact, a defence of inequality and exploitation.

Both in the *Inquiry* and *Labor Rewarded*, therefore, Thompson looked to the transcendence of the market through the creation of co-operative communities. By abolishing exchange these communities eliminated any possibility of exploitation and the material and moral evils which resulted from it. Their neo-autarkic nature insulated them against the vagaries of contemporary capitalism while the distribution of produce according to need freed communitarians from the fear of destitution and from the social antagonism and immorality consequent upon the individual pursuit of material gain. They were, in effect, the ideal environment for the transformation of the human base metal of the old moral order into the gold of the new moral world; a transformation which Thompson hoped to accelerate with the publication in 1830 of his *Practical Directions for the Speedy and Economical Establishment of Communities*.

The most powerful initial influence upon Thompson was Benthamite utilitarianism. This is apparent in the *Inquiry* where Thompson discussed its egalitarian implications and where he considered at length the whole problem of reconciling security with equality. However, the most profound and lasting intellectual influence was that of Owenism, though it should be stressed that Thompson helped to mould Owenite thinking as much as Owenite socialism shaped his own thought. In particular his political or, as Thompson would have preferred it, 'social economy', provided Owenites with a new range of critical tools of analysis with which to condemn the existing order and to sap the ideological defences thrown up by the popularizers of classical economics. His *Inquiry* was the *magnum opus* of co-operative political economy, far surpassing in scope and coherence anything emanating from the pen of Robert Owen.

Thompson died in 1833 and while his *Inquiry* was reprinted in 1850, its influence faded in the second half of the 19th century as co-operation assumed a commercial rather than a communitarian form.

See Also

▶ Ricardian Socialists

Selected Works

1824. *An inquiry into the principles of the distribution of wealth most conductive to human happiness*. London.
1827. *Labour rewarded. The claims of labor and capital conciliated, by one of the idle classes*. London.
1830. *Practical directions for the speedy and economical establishment of communities*. London.

Bibliography

Beales, H.L. 1933. *The early English socialists*. London: Hamish Hamilton.

Beer, M. 1953. *A history of British socialism*, 2 vols. London: Allen & Unwin.
Cole, G.D.H. 1977. *A history of socialist thought*, 5 vols. Vol. 1: *Socialist thought, the forerunners*. London: Macmillan.
Foxwell, H.S. 1899. Introduction to the English translation of A. Menger. In *The right to the whole produce of labour*. London: Macmillan.
Gray, A. 1967. *The socialist tradition, Moses to Lenin*. London: Longman.
Hunt, E.K. 1979. Utilitarianism and the labour theory of value. *History of Political Economy* 11: 544–571.
Hunt, E.K. 1980. The relation of the Ricardian socialists to Ricardo and Marx. *Science and Society* 44: 177–198.
Jones, G.S. 1983. Rethinking chartism. In *Languages of class: Studies in English working class history 1832–1892*. Cambridge: Cambridge University Press.
King, J.E. 1981. Perish commerce! Free trade and underconsumption in early British radical economics. *Australian Economic Papers* 20: 235–257.
Lowenthal, E. 1911. *The Ricardian socialists*. New York: Longman.
Pankhurst, R.K.P. 1954. *William Thompson, pioneer socialist, feminist and co-operator*. London: Watts.
Thompson, N.W. 1984. *The people's science: The popular political economy of exploitation and crisis, 1816–34, 1816–34*. Cambridge: Cambridge University Press.

Thornton, Henry (1760–1815)

David Laidler

Keywords

Bank of England; Bullion Committee; Bullionist controversies; Central banking; Classical theory of money; Convertibility; Credit; Fisher, I.; Forced saving; Hume, D.; Inflation; Lender of last resort; Monetary economics; Neutrality of money; Quantity theory of money; Real bills doctrine; Ricardo, D.; Specie; Thornton, H.; Transfer problem; Velocity of circulation; Wicksell, J. G. K

JEL Classifications

B31

Henry Thornton was born in 1760, the youngest son of John Thornton, a London merchant prominent in the Russian trade. All three of John Thornton's sons were important in the business community and all three served as Members of Parliament. The eldest, Samuel, followed his father in the Russian trade, was director of the Bank of England, and its Governor between 1799 and 1801; Robert served as Governor of the East India Company for a time, but business reverses were eventually to lead to his emigration to the United States; and Henry become an extremely successful London banker. He died, probably from consumption, in 1815.

John Thornton had been and early member of the Evangelicals, as those followers of John Wesley who remained within the Church of England were called, and Henry too was among their leaders, the most famous of whom was his second cousin and close friend William Wilberforce. The movement became known as the Clapham Sect largely because their informal headquarters was Thornton's country house, located in that then outlying village. The Evangelical were also known as 'the Party of Saints' and what we would now regard as the conventional piety and respectability of the Victorian middle classes owe much to their influence. Nevertheless, their milieu was not Victorian but Georgian and Regency England, where their insistence that public policy be informed by the same high moral purpose as their private lives was profoundly radical. Their best-known accomplishment was ending Britain's participation in the slave trade in 1809, and in 1833 the abolition of slavery itself in the British Empire; but the role of their Sunday School Movement in promoting popular literacy in Britain, not to mention the influence of their British and Foreign Bible Society on 19th-century missionary activity throughout the world, is also noteworthy.

Henry Thornton was at the centre of all of these activities and many others as organizer, fund-raiser and donor. Before his marriage in 1796 he habitually devoted six-sevenths of his considerable income to charity, and perhaps a quarter thereafter. During his 33 years' service in Parliament, in addition to his work against the slave trade, he supported such progressive causes as peace with the American colonies,

accommodation with France, and Catholic Emancipation. He also devoted considerable time and energy to religious writings and his great-great-grandson E.M. Forster (1951) records that his posthumously published volume of *Family Prayers* was something of a Victorian bestseller which was still earning royalties for his descendants at the end of the 19th century.

Among all of this activity, Henry Thornton found time to study monetary economics. As a prominent banker and Member of Parliament it was natural that he would take a practical interest in such matters, particularly given the financial turbulence associated with the French Wars of 1793–1815 and the suspension of the gold convertibility of Bank of England notes which accompanied them. He gave evidence to the Parliamentary Committees enquiring into the circumstances of the suspension in 1797, and he was an important member of both the Commons Committee which investigated Irish currency problems in 1804 and the famous 'Bullion Committee' of 1810. However, he was also and above all a great monetary theorist, and his outstanding treatise, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802), gives him a strong claim to be regarded as the most important contributor to monetary economics between David Hume (1752) and Knut Wicksell (1898). Only David Ricardo could seriously be regarded as his rival here.

The early 18th century had seen considerable progress in monetary economics, and David Hume's three essays of 1752 are rightly regarded as containing the core of classical monetary theory. They set out the quantity theory doctrine that, other things being equal, the price level varies with the quantity of money, and accompany this with an analysis of the way in which, under a commodity standard, balance of payments mechanisms operate so as to equalize price levels and distribute the precious metals among countries. Although allowing that monetary changes can have short-run effects on real output, they also develop the basic classical postulate that money is neutral in the long run, affecting only prices; and in particular they argue that the rate of interest is not a monetary phenomenon.

Banks are scarcely mentioned in Hume's analysis, and though Adam Smith (1776) paid considerable attention to them, his model was the 18th-century Scottish system. Scottish commercial banks held their reserves in claims upon London, not upon any Scottish central bank, and Scotland was a small, largely price-taking economy. Hence Smith's analysis of the interaction of bank behaviour, the price level and the balance of payments, though remarkably perceptive, was far from complete. It had little to say about the transmission mechanisms at work here and about the role of financial assets other than banknotes in the monetary system. Moreover it had nothing at all to say about central banking.

By the 1790s, the development of the English monetary system had far outstripped the growth of knowledge concerning the principles that underlay its operations, and the financial crisis which culminated in the suspension of February 1797 drew attention to this gap in most dramatic fashion. Thornton's *Paper Credit,* published in 1802 but perhaps begun as early as 1796, not only remedied this deficiency, but brought monetary theory to a level of sophistication that it was not to surpass until the end of the 19th century, as a brief sketch of its contributions will make quite evident.

Paper Credit begins with a detailed description of the contemporary English monetary system, showing how a rather wide variety of credit instruments had come to circulate as what we would now call money, alongside coin and banknotes, and it argues that the velocities of various components of this complex 'circulating medium' differ among instruments and fluctuate over time. In common with virtually every monetary economist before Irving Fisher (and many thereafter), Thornton regarded velocities of circulation as frequently unstable and he discussed in some detail how the Bank of England should behave, both to minimize the occurrence of monetary instability and to offset its consequences when it arose. Thornton was by no means the only contributor to the 'Bullionist Controversy', as the debates of the period are called, to recognize the crucial role and responsibilities of the Bank of England as a central bank, but there were many, not least among the directors

of that institution, who refused to do so; and Thornton's exposition of the issues involved represents an important contribution to monetary economics.

No doubt drawing upon his own first-hand observations of the mechanisms at work during the turbulent 1790s, Thornton stressed both the crucial role and the volatility of the public's confidence in the banking system's ability to redeem its liabilities (in terms of Bank of England notes in the case of country and private London banks, and, under convertibility, in terms of specie in the case of the Bank of England). He understood that bank customers, who were confident that they could obtain Bank of England notes or specie when they required it, would not in fact seek such accommodation, and that only those who had doubts about the convertibility of their assets would demand their redemption. Hence he argued that any initial fall in confidence could lead to a self-reinforcing drain of reserves from the system if the Bank of England responded to it by reducing lending and hence cutting down the supply of the very central bank notes that the public were demanding from country and London banks. For Thornton, the right response to such an 'internal' (that is, within the country) drain of reserves from the banking system was for the Bank of England to lend freely to all solvent borrowers in order to restore and maintain the public confidence in the system. In short, the by now conventional textbook analysis of the central bank's 'lender of last resort' function found its first full statement in *Paper Credit.*

But Thornton understood well enough that an internal drain was not the only possible source of pressure on reserves. An external drain associated with what we would now call an adverse balance of payments was also a possibility, and here the required remedy might be different. He was clear that, to the extent that the drain stemmed from an uncompetitively high domestic price level, it could only be remedied by monetary contraction, and hence by the central bank scaling down its loans, including those made to the rest of the banking system. In the conventional wisdom of the later 19th century concerning sound central bank practice, an external drain was always appropriately to be met by such measures, but Thornton (unlike Ricardo, who is the true father of that conventional wisdom) was more subtle than this in his analysis.

For him, money wages were sticky and any sudden monetary contraction carried with it the danger of disrupting markets and causing real output and employment to fall, a danger to be avoided if at all possible. Hence when developing the implications for Bank of England policy of his pioneering analysis of what was later to be called 'the transfer problem', he advocated that temporary drains of specie abroad, associated with bad harvests or once and for all subsidy payments to allies, be accompanied by as little domestic monetary contraction as seemed to that institution to be prudent. Under arrangements prevailing after 1797, he was even willing to entertain temporary departures of sterling from par with specie in the face of temporary external drains rather than risk the domestic disruption that might accompany monetary contraction. Thornton was thus in *Paper Credit* far from being an advocate of an automatic gold standard, and his views have something in common with those of such later advocates of managed paper currency as Thomas Attwood – not to mention John Maynard Keynes, as certain commentators, notably Hicks (1967) and Beaugrand (1981) have pointed out.

McCulloch (1845), who confused Henry with his brother Samuel, regarded *Paper Credit* as being too partial to the Bank of England in its arguments, but though it may certainly be regarded as a defence of that institution's behaviour during the early years of the restriction, it is nevertheless a critical defence. Even so, by 1810, Henry Thornton was a prominent member of the Bullion Committee, and had become one of the Bank's sternest critics, advocating, both as a signatory to the Committee's Report (Cannan 1919) and in two Commons speeches on the Report, that the obligation to redeem its notes in specie be reimposed upon it as soon as possible, a measure which was designed to narrow considerably the scope for discretion left to the Bank when confronted with an external drain.

Thornton's policy stance had changed between 1802 and 1810, but there is no evidence that his underlying analytic views were any different. First and foremost, and despite certain affinities,

mentioned above, between his work and that of subsequent advocates of managed paper standards, Henry Thornton was always, as Hicks (1967) has put it, a 'hard money' man as far as long-run policy questions were concerned. He regarded the maintenance of the specie value of Bank of England liabilities as the proper overriding end of monetary policy. After 1797 he expected the Bank of England, subject to certain caveats about bad harvests and once and for all transfers, to manage its discounts so as to stabilize the exchange rate and the price of specie. In 1802 he believed that the Bank could be trusted to do so without the check of convertibility, but by 1810 he had changed his mind.

Though the actual conduct of monetary policy, particularly after 1811, shows that, luckily for Britain, they did not always practise what they preached, the directors of the Bank declared themselves firmly committed to the so-called 'real bills doctrine' in their evidence to the Bullion Committee, as they did in many other statements. This doctrine distinguishes between 'real bills', drawn to finance goods in the process of production and distribution, and 'fictitious bills' those which simply represent a debt with no corresponding real asset to back them. It then argues that a banking system in general, and a central bank in particular, which confines its activities to the discount of the former, cannot affect the price level. The quantity of money generated by following such practices will, so it is claimed, vary with the volume of output and adjust itself automatically and passively to the 'needs of trade'.

Thornton had considered and comprehensively refuted this bundle of fallacies in *Paper Credit*. He had shown that, because there is no necessary relationship between the period for which commercial bills are discounted and the period of time that elapses between the beginning of the production of a particular unit of output and its final consumption, the distinction between 'real' and 'fictitious' bills was specious. Distinguishing between credit per se, and the role of credit instruments as components of the circulating medium, he had also shown how money, even if created against the security of good quality commercial bills, could influence the price level. Finally, and crucially, he had shown that the demand of manufacturers and merchants for bank credit would vary with the relationship between the banking system's lending rate and the expected rate of profit in such a way that, if the latter were high relative to the former, potentially unlimited monetary expansion and inflation could be generated by a banking system whose central authority took the real bills doctrine as its sole operating guide.

These arguments of Thornton's play a central role in the 1810 *Report* of the Bullion Committee and reflect his influence on that document. The explicit rejection of them by the directors of the Bank of England, not to mention widespread concern about inflation during 1809–10, was a crucial factor in persuading the committee in general, and Thornton in particular as one of its key members, to recommend that the constraint of specie convertibility be reimposed upon the Bank as soon as possible, a recommendation which was, of course, rejected by Parliament in 1811 along with the rest of the *Bullion Report*.

The reader familiar with the later literature of monetary economics will recognize the essentially Wicksellian (for example, 1898) flavour of Thornton's discussion of the relationship between bank lending policies and inflation. In a parliamentary speech of 1811 on the *Bullion Report* he elaborated on his earlier analysis by allowing for the influence of inflation expectations on the perceived real interest burden implied by any given nominal bank lending rate. This insight, which plays only an occasional and peripheral role in Wicksell's work, was of course central to the contributions of another great monetary theorist, Irving Fisher (1896). Moreover in his analysis of these matters, Thornton developed a version of what was later to be called the 'forced saving' doctrine which played an important role in early 20th-century business cycle theory. In the light of all this, it would be easy to jump to the conclusion that Thornton's work was well known to his successors. However, it was not.

Failing health and a relatively early death removed Henry Thornton from the centre of monetary controversy just as David Ricardo came to the height of his powers and influence. It was Ricardo and not Thornton who was destined to

become the recognized authority to whom 19th-century monetary economists working within the classical tradition looked for guidance in matters of monetary theory. As Hutchison (1968) points out, J.S. Mill (1848) was the last important 19th-century author to recognize Thornton's contributions. Thereafter, his name faded from view, and was not even known to Wicksell; it is largely due to the efforts of Jacob Viner (1924, 1937) and particularly Friedrich von Hayek (1939) that his true stature came to be appreciated in the 20th century. Nevertheless, his ideas were well known to his contemporaries, not least to Ricardo, and as transmitted by them, not always without a certain loss of subtlety, they permeate 19th-century classical monetary theory. Thus if Henry Thornton's name was often forgotten by economists, his contributions to the subject were certainly not. For most men, this would be small consolation indeed, but one suspects that so benevolent and self-effacing a man as Henry Thornton might have been content with such an outcome.

See Also

- ▶ Hume, David (1711–1776)
- ▶ Quantity Theory of Money

Selected Works

1802. *An enquiry into the nature and effects of the paper credit of Great Britain.* Together with his evidence given before the Committees of Secrecy of the two Houses of Parliament in the Bank of England, March and April 1797, some manuscript notes, and his speeches on the Bullion Report, May 1811. Edited with an introduction by F.A. von Hayek, London: George Allen & Unwin, 1939. Reprinted, London: Frank Cass & Co.; New York: Augustus Kelley, 1962.

Bibliography

Beaugrand, P. 1981. *Henry Thornton: Un precurseur de J.M. Keynes*. Paris: Presses Universitaires de France.

Cannan, E. 1919. *The paper pound of 1797–1821: The bullion report*. London: P.S. King & Son. 2nd edn, 1921, reprinted New York: Augustus M. Kelley, 1969.

Fisher, I. 1896. Appreciation and interest. *AEA Publications* 3 (11): 331–442.

Forster, E.M. 1951. Henry Thornton. In *Two cheers for democracy*, ed. E.M. Forster. London: Abinger. Reprinted, London: Edward Arnold, 1972.

Hicks, J.R. 1967. Thornton's *paper credit*. In *Critical essays in monetary theory*. London: Oxford University Press.

Hume, D. 1752. Of money; Of the balance of trade; Of interest. In *Essays, moral, political and literary,* ed. D. Hume. London. Reprinted, London: Oxford University Press, 1962.

Hutchison, T.W. 1968. Thornton, Henry. In *International encyclopaedia of the social sciences*, vol. 16. New York: Macmillan and Free Press.

McCulloch, J.R. 1845. *The literature of political economy.* London.

Mill, J.S. 1848. *Principles of political economy with some of their applications to social philosophy.* London: Parker.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations.* 2 vols, ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Clarendon Press, 1976.

Viner, J. 1924. *Canada's balance of international indebtedness 1900–1913*. Cambridge, MA: Harvard University Press.

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper Bros.

von Hayek, F.A. 1939. Introduction to Thornton (1802).

Wicksell, K. 1898. *Interest and prices.* Trans. R.F. Kahn. London: Macmillan for the Royal Economic Society, 1936.

Thornton, William Thomas (1813–1880)

A. Picchio

Thornton was born in Buckinghamshire and lived three years in Malta and five years in Constantinople. In 1836 he obtained a clerkship in the East India House and later (1858) became secretary for public works to the India Office. His expertise had a wide range, from literary works to political economy. Thornton's economic works were praised by his friend J.S. Mill, who referred to his work on population in the *Principles* and used

Thornton's arguments for his recantation of the strict wages-fund theory (Mill 1869). Thornton's fortunes with Marshall were less stable. Marshall was doubtful about the critique of the law of supply and demand as the determinant of wages but appreciated Thornton's style and his work on trade unions (Marshall 1975, pp. 117–20, 263; 1960, p. 365).

In 1846 Thornton published a book on population based on the wages-fund theory, and then in 1869 he published *On Labour* presenting his critique. The latter book appeared after Longe's contribution on the same topics, and followed Longe's reasoning. Longe's originality was recognized by the Political Economy Society (Hollander 1903, p. 5), and the possibility of Thornton's plagiarism was discussed by various authors (Hollander 1903; Schumpeter 1954, pp. 669–70).

Thornton's critique of the wages fund aimed mainly to show, with many detailed examples, that supply-and-demand adjustments of prices to quantities and vice versa did not occur in the real world. First of all Thornton considered prices as usually 'reserved', hence not flexible in relation to changes in quantities. Even in the case of unreserved prices, quantities would not adjust symmetrically to relative changes in supply and demand. Thornton in fact considered prices and quantities as separately determined. He thought that although it is usually true that prices rise when demand exceeds supply, it is not true that demand decreases if prices increase (Thornton 1870, pp. 58–64). Moreover, prices can remain stable if demand and supply change (p. 67), and, even more devastating for supply-and-demand theories, demand and supply are not brought into equilibrium, as general case, by changes in prices (pp. 73–5).

Thornton does acknowledge the role of competition among buyers and sellers as a determinant of prices, but he does not assume any systematic behaviour which could justify the idea of general laws of prices.

When Thornton moves form his critique of supply-and-demand prices to its application to the determination of wages, his alliance with Long against the dogmatism of theories assuming a mechanical determinacy of wages emerges clearly (Thornton 1970, pp. 82–5; McNulty 1980, p. 80). Thornton sees insecurity as the basis for the specificity of labour as a commodity. Although labour is sold at unreserved price, there is still no mechanism which can be assumed to adjust quantities to prices. For this reason and because the demand for labour cannot be considered a definite fund – as it depends on capitalists' consumption and expectations – wages cannot be considered as determinate within the supply-and-demand allocation of labour (ibid, pp.85–105).

On the question of labour relations Thornton was a great supporter of Cooperation. Cooperation was considered a possible alternative to class war and strikes, and the basis for an alliance between labour and capital. Cooperation was debated in the Unions at the time (Webb [1894] 1920, p. 225) and it reconciled Thornton with his friend J.S. Mill with regard to policies if not to theories.

See Also

- ▶ Mill, John Stuart (1806–1873)
- ▶ Wage Fund Doctrine

Selected Works

1846. *Overpopulation and its remedy; or, an enquiry into the extent and causes of the distress prevailing among the labouring classes of the British Islands*. London: Longman, Brown, Green, & Longmans.

1848. *A plea for peasant proprietors*. London: J. Murray.

1854. *Zohrab; or a midsummer day's dream; and other poems,* London.

1867a. What determines the price of labour or rate of wages? *Fortnightly Review.*

1867b. Stray chapters from a forthcoming work on labour. *Fortnightly Review.*

1869. *On Labour, its wrongful claims and wrightful dues; its actual present and possible future.* London: Macmillan.

1873. *Old fashioned ethics and common-sense metaphysics with some of their applications*. London: Macmillan.
1875. *Indian public works, and cognate Indian topics*. London: Macmillan.
1878. *Horatius Flaccus, word for word from Horace. The Odes literally versified by William Thomas Thornton*. London.

Bibliography

Hollander, J.H. 1903. Introduction to F.D. Longe. In *The wages fund theory*. Baltimore: Johns Hopkins Press.
McNulty, P.J. 1980. *The origins and development of labor economics*. Cambridge, MA: Harvard University Press.
Marshall, A. 1975. In *The early economic writings of Alfred Marshall, 1867–1890*, vol. I and II, ed. J.K. Whitaker. London: Macmillan.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Webb, S., and B. Webb. 1894. *The history of trade unionism*. London: Longman & Co. Revised ed, 1920.

Threshold Models

Timo Teräsvirta

Abstract

This article contains a short account of threshold, smooth transition and Markov switching autoregressive models. Neural network models are highlighted as well. Linearity testing, parameter estimation and, more generally, modelling are considered. Forecasting with threshold models receives attention. Suggestions for further reading are supplied.

Keywords

ARIMA models; Artificial neural network models; Asymmetric behaviour; Forecasting; Identification; Likelihood; Linear models; Logistic smooth transition regression models; Markov chains; Markov-switching models; Maximum likelihood; Modelling; Purchasing power parity; Regression error specification test; Smooth transition regression models; Switching regression models; Threshold autoregressive models; Threshold models

JEL Classification

C10

The Models

Stochastic nonlinear models have been widely used in economic applications. They may arise directly from economic theory. There also exist nonlinear models that have first been suggested by statisticians, engineers and time series analysts and then found application in economics. A broad class of these models, here called threshold models, has the property that the models are either piecewise linear or may be more generally considered as linear models with time-varying parameters. This category of nonlinear models includes switching regression or threshold autoregressive models, smooth transition models, Markov-switching or hidden Markov models. Artificial neural network models may also be included in this class of nonlinear models.

The switching regression (SR) model or, in its univariate form, the threshold autoregressive (TAR) model, is defined as follows:

$$y_t = \sum_{j=1}^{r} \left(\alpha_j' \mathbf{z}_t + \varepsilon_{jt}\right) I\left(c_{j-1} < s_t \leq c_j\right) \quad (1)$$

where $\mathbf{z}_t = (\mathbf{w}'_t, \mathbf{x}'_t)'$ is a vector of explanatory variables, $\mathbf{w}_t = (1, y_{t-1},\ldots, y_{t-p})'$ and $\mathbf{x}_t = (x_{1t},\ldots, x_{kt})'$, s_t is an observable switch-variable, usually assumed to be a continuous stationary random variable, $c_0, c_1,\ldots, c_r$ are switch or threshold parameters, $c_0 = -\infty$, $c_r = M < \infty$, and $I(A)$ is an indicator function: $I(A) = 1$ when event A occurs; zero otherwise. Furthermore, $\alpha_j = (\alpha_{0j}, \alpha_{1j},\ldots, \alpha_{mj})'$ such that $\alpha_i \neq \alpha_j$ for $i \neq j$, where $m = p + k + 1$, $\varepsilon_{jt} = \sigma_j \varepsilon_t$ with $\{\varepsilon_t\} \sim$ iid (0, 1), and $\sigma_j > 0, j = 1,\ldots, r$. It is seen that (1) is a piecewise linear model whose switch-points, however, are generally unknown. The most popular choice in applications is $r = 2$, that is, the

model has two regimes. If $s_t = t$, Eq. (1) is a linear model with $r - 1$ breaks and $\alpha_j \neq \alpha_{j+1}, j = 1,\ldots, r - 1$. These models have recently become quite popular in econometrics and there is now a substantial literature on how to determine the number of breaks and estimate the break points. For a generalization of the threshold autoregressive regression model to the vector case, see Tsay (1998).

TAR models have been used to characterize asymmetric behaviour in GNP or unemployment rates and to consider the purchasing power parity hypothesis. They have also been applied to modelling interest rate series as well as other financial time series.

One may substitute $I\ (s_t = j)$ for $I\ (c_{j-1} < s_t \leq c_j)$ in (1), where s_t is an unobservable regime indicator with a finite set of values $\{1,\ldots, r\}$. On the assumption that s_t follows a first-order Markov chain, that is, $\Pr\{s_t = i|s_t = j\} = p_{ij},\ i, j = 1,\ldots, r$, (1) becomes a hidden Markov or Markov-switching (MS) model (see Lindgren 1978). Higher-order Markov chains are possible but rarely used in econometric applications.

Equation (1) with an unobservable regime indicator is not, however, the most frequently applied hidden Markov model in econometrics. Consider the univariate model

$$y_t = \mu_{s_t} + \sum_{j=1}^{p} \alpha_j \left(y_{t-j} - \mu_{s_{t-j}} \right) + \varepsilon_t \qquad (2)$$

where s_t follows a first-order Markov chain as before, and $\mu_{s_t} = \mu^{(i)}$ for $s_t = i$, such that $\mu^{(i)} \neq \mu^{(j)}, i \neq j$. The stochastic intercept of this model, $\mu_{s_t} - \sum_{j=1}^{p} \alpha_j \mu_{s_{t-j}}$, can thus obtain r^{p+1} different values, which gives the model the desired flexibility. A comprehensive discussion of MS models can be found in Hamilton (1994, ch. 22). The model (2) has been frequently fitted, for example, to GNP series and interest rates. In the latter case, the model may be used for identifying changes in the monetary policy of the central bank.

SR and Markov-switching models contain a finite number of regimes. There is a class of models called smooth transition regression (STR) models, in which there are two 'extreme regimes', and the transition between them is smooth. A basic STR model is defined as follows:

$$y_t = \varphi' \mathbf{z}_t + \psi' z_t\, G(\gamma, \mathbf{c}, s_t) + \varepsilon_t \qquad (3)$$

where $\varphi = (\varphi_0, \varphi_1,\ldots, \varphi_m)'$ and $\psi = (\psi_0, \psi_1,\ldots, \psi_m)'$ are parameter vectors, $\mathbf{c} = (c_1,\ldots, c_K)'$ is a vector of location parameters, $c_1 \leq \ldots \leq c_K$, and $\varepsilon_t \sim \text{iid}(0, \sigma^2)$. The transition function $G(\gamma,\mathbf{c},s_t)$ is a bounded function of s_t, continuous everywhere in the parameter space for any value of s_t. The logistic transition function has the general form

$$G(\gamma, c, s_t) = \left(1 + \exp\left\{ -\gamma \prod_{k=1}^{K} (s_t - c_k) \right\} \right)^{-1}, \quad \gamma > 0 \qquad (4)$$

where $\gamma > 0$ is an identifying restriction. Equation (3) jointly with (4) defines the logistic STR (LSTR) model. The most common choices in practice for K are $K = 1$ and $K = 2$. For $K = 1$, the parameters $\varphi + \psi G(\gamma, \mathrm{c}, s_t)$ change monotonically as a function of s_t from φ to $\varphi + \psi$. For $K = 2$, they change symmetrically around the mid-point $(c_1 + c_2)/2$ where this logistic function attains its minimum value. Slope parameter γ controls the slope and c_1 and c_2 the location of the transition function. When $K = 1$ and $\gamma \to \infty$ in (4), the model (3) becomes an SR model with $r = 2$.

The LSTR model with $K = 1$ (LSTR1 model) is capable of characterizing asymmetric behaviour. As an example, suppose that s_t measures the phase of the business cycle. Then the LSTR1 model can describe processes whose dynamic properties are different in expansions from what they are in recessions, and the transition from one extreme regime to the other is smooth. The same is true for SR and the MS models with the difference that instead of a smooth transition there is an abrupt switch. The LSTR2 model is appropriate whenever the dynamic behaviour of the process is similar at both large and small values of s_t and different in the middle.

Yet another nonlinear model that is worth mentioning because it is related to threshold models is the artificial neural network (ANN) model.

T

The simplest single-equation case is the so-called 'single hidden-layer' ANN model. It has the following form

$$y_t = \beta_0' \mathbf{z}_t + \sum_{j=1}^{q} \beta_j G\left(\gamma_j' \mathbf{z}_t\right) + \varepsilon_t \qquad (5)$$

where y_t is the output series, z_t is the vector of inputs, and $\beta_0' \mathbf{z}_t$ is a linear unit with $\beta_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0,p+k})'$. Furthermore, $\beta_j, j = 1, \ldots, q$, are parameters, called connection strengths" in the neural network literature. Function G(·) is a bounded, asymptotically constant function called the 'squashing function' and $\gamma_j, j = 1, \ldots, q$, are parameter vectors. They form the hidden layer which the name of the model refers to. Typical squashing functions are monotonically increasing ones such as the logistic function and the hyperbolic tangent function. The errors ε_t are often assumed iid $(0,\sigma^2)$. Many neural network modellers assume $\beta_0 = \beta_{00}$, where β_{00} is called the 'bias'.

A theoretical argument for the use of ANN models is that they are universal approximators. Suppose that $y_t = H(\mathbf{z}_t)$, that is, there exists a functional relationship between y_t and $\mathbf{z}_t$. Then, under mild regularity conditions for H, there is a positive integer $q \leq q_0 < \infty$ such that for an arbitrary $\delta > 0$, $\left|H(\mathbf{z}_t) - \sum_{j=1}^{q} \beta_j G\left(\gamma_j' \mathbf{z}_t\right)\right| < \delta$. The importance of this result lies in the fact that q is finite, so that any unknown function H can be approximated arbitrarily accurately by a linear combination of squashing functions $G(\gamma' z_t)$. This has been discussed in several papers, including Cybenko (1989), Funahashi (1989) and Hornik et al. (1989). Neural network models are very generously parameterized and are only locally identified. The log-likelihood typically contains a large amount of local maxima, which makes parameter estimation difficult.

Testing Linearity

All threshold models nest a linear model but they are not identified when the data are generated from this linear model. For this reason, testing linearity before fitting a threshold model is necessary in order to avoid the estimation of an unidentified model whose parameters cannot be estimated consistently. In this case, linearity testing has to precede any nonlinear estimation.

There exist general misspecification tests that are linearity tests if the specification to be tested is linear; see Bierens (1990) and Stinchcombe and White (1998). There also exist parametric tests that have been designed to be tests against an unspecified alternative but are not consistent against deviations from linearity. The popular Regression Error Specification Test (RESET) of Ramsey (1969) is such a test. Teräsvirta (1998) and van Dijk et al. (2002) discuss tests against smooth transition regression models and Hansen (1999) surveys linearity testing against TAR models. Linearity testing in the Markov-switching framework is considered in Garcia (1998). Some recent econometrics textbooks discuss linearity tests against various threshold models.

As already mentioned, threshold models nest a linear model and are not identified if linearity holds. For example, the STR model (3) is not identified if $\gamma = 0$ in (4) or $\psi = 0$. In the former case, ψ and c are not identified, and in the latter, the nuisance parameters are γ and c. Consequently, the standard asymptotic theory is not applicable in testing linearity. This problem may be solved following Davies (1977). Let γ be the vector of nuisance parameters. For example, $\gamma = (\gamma, c')'$ in (3) when the null hypothesis is $\psi = \mathbf{0}$. When γ is known testing linearity is straightforward. Let $S_T(\gamma)$ be the corresponding test statistic whose large values are critical and define $\Gamma = \{\gamma : \gamma \in \Gamma\}$, the set of admissible values of γ. When γ is unknown, the statistic is not operational because it is a function of γ. The problem is solved by defining another statistic $S_T = sup_{\gamma \in \Gamma} S_T(\gamma)$ that is free of nuisance parameters γ. The asymptotic distribution of S_T under the null hypothesis does not generally have an analytic form, but Davies (1977) gives an approximation to it that holds under certain conditions, including the assumption that $S(\gamma) = plim_{T\to\infty} S_T(\gamma)$ has a derivative. Other choices of test statistic include the average:

$$S_T = \text{ave} S_T(\gamma) = \int_\Gamma S_T(\gamma) dW(\gamma) \qquad (6)$$

where $W(\gamma)$ is a weight function defined by the user such that $\int_\Gamma W(\gamma)\mathrm{d}\gamma = 1$, and the exponential

$$\exp S_T = \ln\left(\int_\Gamma \exp\{(1/2)S_T(\gamma)\} dW(\gamma)\right). \qquad (7)$$

Andrews and Ploberger (1994) have recommended these tests and demonstrated their local asymptotic optimality properties. The statistics (6) and (7) are two special cases in the family of average exponential tests (for definitions and details, see Andrews and Ploberger 1994). Hansen (1996) shows how to obtain asymptotic critical values for these statistics by simulation under rather general conditions. His method is computationally intensive but useful. It may be pointed out that it works for SR and STR models where s_t is observable. For MS models, the situation is more complicated, see Garcia (1998) for discussion.

A computationally simpler alternative is to circumvent the identification problem instead of directly solving it. It has been popular in testing linearity against smooth transition models, eq. (3). The idea is to replace the transition function (4) by its Taylor series approximation around the null hypothesis $\gamma = 0$. This transforms the testing problem into one of testing a linear hypothesis in a linear auxiliary regression, see Luukkonen et al. (1988) or Teräsvirta (1998). Parameter estimation Parameters of threshold models have to be estimated numerically. This is because the objective function to be optimized is not quadratic in parameters, so an analytical solution to the problem does not exist. The easiest models to estimate are the switching regression or threshold autoregressive models. Their parameters are estimated conditionally by ordinary least squares (OLS), given the switch parameters $c_1,\ldots,c_r$, and the combination of $c_1,\ldots,c_r$ yielding the smallest sum of squared residuals gives the estimates of these and the other parameters. For example, when $r = 1$ the OLS estimation is repeated for a set of c_1 values such that both regimes contain at least a certain minimum amount of observations, typically 10% or 15% of the total number. Under rather general conditions, including stationarity and ergodicity of the TAR process, the least squares estimators are $\sqrt{T}$ -consistent and asymptotically normal. The threshold parameter estimators are super (T-) consistent.

Smooth transition models are estimated using standard maximum likelihood. The most efficient numerical method is the Newton–Raphson method that makes use of both the first (the score) and the second (the Hessian) partial derivatives of the log-likelihood function. It has many variants in which the Hessian is replaced by computationally simpler alternatives that either do not require second derivatives, such as the method of scoring or the so-called Berndt–Hall–Hall–Hausman (BHHH) algorithm, or avoid inverting the Hessian altogether. Examples of this include the steepest descent and variable metric methods. Of the latter, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm can be found in a number of modern software packages.

Hidden Markov models cannot be estimated using standard optimization algorithms because they contain the latent variable s_t. Their parameters are typically estimated using the expectation–maximization (EM) algorithm (see Cappé et al. 2005, and Hamilton 1994, ch. 22).

Estimation of ANN models using maximum likelihood is often numerically demanding because the likelihood function can contain a large number of local maxima, due to a large number of parameters. This problem has been discussed in Medeiros et al. (2006) and White (2006). Because of this difficulty, the literature on ANN models contains a wide variety of estimation methods of various kinds; see for example Fine (1999) and White (2006).

Modelling

Typically, economic theory does not uniquely determine the functional form of a threshold model. This means that the model builder has to select an appropriate model for the problem at hand. In this case, applying a consistent modelling strategy is helpful. The modelling approach of

Box and Jenkins (1970) for the ARIMA class of linear models is a case in point. When it comes to threshold models, modelling strategies consisting of stages of specification, estimation and evaluation have been worked out and applied for TAR or, more generally, SR classes of models as well as STR models. For the former, see Tsay (1989) (univariate models) and Tsay (1998) (multivariate models) and for the latter, Teräsvirta (1998) or Teräsvirta (2004). Medeiros et al. (2006) suggest a similar procedure for ANN models. An essential first stage in all these strategies is testing linearity. If linearity is not rejected, the task of the model builder is considerably simplified.

Forecasting

The main purpose of univariate nonlinear models is forecasting. Multivariate models may also be useful for policy analysis. Forecasts are typically conditional means in which the conditioning set consists of a subset of the information available at the time of making the forecast. In nonlinear models such as threshold models, a typical situation is that making forecasts for more than one period ahead requires numerical methods. This is due to the fact that for a random variable X, generally $Eg(X) \neq g(EX)$. Equality holds if g is a linear function of X.

To illustrate, assume an information set $\mathcal{F}_T$ at time T. The optimal one-period mean forecast $f^y_{T,1} = E\{y_{T+1}|\mathcal{F}_T\}$, the conditional mean of y, given $\mathcal{F}_T$. For example, consider the simple bivariate model

$$y_t = g(x_{t-1}) + \varepsilon_t \tag{8}$$

where

$$x_t = \beta x_{t-1} + \eta_t \tag{9}$$

with $|\beta| < 1$, and $\{\eta_t\}$ is a sequence of independent, identically distributed random variables with zero mean. Function $g(\cdot)$ may define an SR, STR or ANN model. The forecast for y_{T+1} equals $f^y_{T,1} = g(x_T)$ as $E\{\varepsilon_{T+1}|F_T\} = 0$. Thus, if one knows the function $g(\cdot)$, one-step forecasts can be obtained with no difficulty.

The optimum two-step forecast is

$$f^y_{T,2} = E\{y_{T+2}|\mathcal{F}_T\} = E\{g(x_{T+1})|\mathcal{F}_T\}.$$

As x_{T+1} is not usually known at time T, it has to be forecast from its autoregressive equation. This gives a one-step OLS forecast $f^y_{T,1} = \beta x_T$. The two-step forecast equals

$$f^y_{T,2} = E\left\{g\left(f^y_{T,1} + \eta_{T+1}\right)|\mathcal{F}_T\right\} \tag{10}$$

The exact forecast equals

$$fe^y_{T,2} = \int_{-\infty}^{\infty} g\left(\overset{y}{T},_1 + z\right)\mathrm{d}D(z)$$

where $D(z)$ is the cumulative distribution function of z. The integral has to be calculated numerically. It may, however, also be approximated by simulation or by bootstrapping the residuals of the estimated model; see Granger and Teräsvirta (1993) or Teräsvirta (2006a). This alternative becomes even more practical when the forecast horizon exceeds two periods. Yet another alternative is to ignore the error η_{T+1}, but the "naive" forecast $fn^y_{T,2} = g\left(f^x_{T,1}\right)$ is biased. In practice, the function $g(\cdot)$ is not known and has to be specified and estimated from the data before forecasting. One may also obtain the forecast directly as

$$fd^y_{T,2} = E\left(y_{T+2}|\mathcal{F}_T\right)$$

so that $y_{t+2} = g_2(x_t, y_t) + \varepsilon^*_t$, say, and the function $g_2(\cdot)$ has to be determined and estimated separately, rather than derived from the one-step representation (8). A difficulty with this approach is that the errors will not necessarily be white noise. A separate forecast function is needed for each forecast horizon.

All forecasts from hidden Markov models can be obtained analytically by a sequence of linear operations. This is a direct consequence of the fact that the regimes in (1) when $I\ (c_{j\text{-}1} < s_t \leq c_j)$ is

replaced by $I(s_t = j)$, where s_t is a latent discrete variable, are linear in parameters. This is discussed for example in Hamilton (1993) or Teräsvirta (2006a).

Experiences from large empirical studies in which macroeconomic variables are forecast with threshold models, are mixed. No model dominates the others, and in several cases nonlinear threshold models do not improve the accuracy of point forecasts compared to linear models. Recent studies of this type include Stock and Watson (1999), Marcellino (2004) and Teräsvirta et al. (2005).

Further Reading

Many statistics and econometrics monographs contain accounts of threshold models, among them Franses and van Dijk (2000), Granger and Teräsvirta (1993) and Guégan (1994). Tong (1990) focuses on TAR models. There are also useful book chapters and review articles such as Bauwens et al. (2000) offering a Bayesian perspective, Brock and Potter (1993), Teräsvirta (2006a,b) and Tsay (2002). For hidden Markov models, see Cappé et al. (2005) and Hamilton (1993, 1994, ch. 22). The latter reference concentrates on the autoregressive model (2). Several thorough treatments of ANN models exist; see for example Fine (1999) or Haykin (1999).

See Also

- ▶ Forecasting
- ▶ Identification
- ▶ Model Selection
- ▶ Non-linear Time Series Analysis
- ▶ Statistical Inference
- ▶ Testing

Bibliography

Andrews, D.W.K., and W. Ploberger. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62: 1383–1414.

Bauwens, L., M. Lubrano, and J.-F. Richard. 2000. *Bayesian inference in dynamic econometric models*. Oxford: Oxford University Press.

Bierens, H.J. 1990. A consistent conditional moment test of functional form. *Econometrica* 58: 1443–1458.

Box, G.E.P., and G.M. Jenkins. 1970. *Time series analysis, forecasting and control*. San Francisco: Holden-Day.

Brock, W.A., and S.M. Potter. 1993. Nonlinear time series and macroeconometrics. In *Handbook of statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao, and H.D. Vinod. Amsterdam: North-Holland.

Cappé, O., E. Moulines, and T. Rydén. 2005. *Inference in Hidden Markov models*. New York: Springer.

Cybenko, G. 1989. Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals, and Systems* 2: 303–314.

Davies, R.B. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64: 247–254.

Fine, T.L. 1999. *Feedforward neural network methodology*. Berlin: Springer.

Franses, P.H., and D. van Dijk. 2000. *Non-linear time series models in empirical finance*. Cambridge: Cambridge University Press.

Funahashi, K. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks* 2: 183–192.

Garcia, R. 1998. Asymptotic null distribution of the likelihood ratio test in Markov switching models. *International Economic Review* 39: 763–788.

Granger, C.W.J., and T. Teräsvirta. 1993. *Modelling nonlinear economic relationships*. Oxford: Oxford University Press.

Guégan, D. 1994. *Séries chronologiques non linéaires à temps discret*. Paris: Economica.

Hamilton, J.D. 1993. Estimation, inference and forecasting of time series subject to changes in regime. In *Handbook of statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao, and H.R. Vinod. Amsterdam: North-Holland.

Hamilton, J.D. 1994. *Time series analysis*. Princeton: Princeton University Press.

Hansen, B.E. 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64: 413–430.

Hansen, B.E. 1999. Testing for linearity. *Journal of Economic Surveys* 13: 551–576.

Haykin, S. 1999. *Neural networks. A comprehensive foundation*, 2nd ed. Upper Saddle River: Prentice-Hall.

Hornik, K., M. Stinchcombe, and H. White. 1989. Multi-layer feedforward networks are universal approximators. *Neural Networks* 2: 359–366.

Lindgren, G. 1978. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics* 5: 81–91.

Luukkonen, R., P. Saikkonen, and T. Teräsvirta. 1988. Testing linearity against smooth transition autoregressive models. *Biometrika* 75: 491–499.

Marcellino, M. 2004. Forecasting EMU macroeconomic variables. *International Journal of Forecasting* 20: 359–372.
Medeiros, M.C., T. Teräsvirta, and G. Rech. 2006. Building neural network models for time series: A statistical approach. *Journal of Forecasting* 25: 49–75.
Ramsey, J.B. 1969. Tests for specification errors in classical least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31: 350–371.
Stinchcombe, M.B., and H. White. 1998. Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* 14: 295–325.
Stock, J.H., and M.W. Watson. 1999. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In *Cointegration, causality and forecasting: A Festschrift in honour of Clive W. J. Granger*, ed. R.F. Engle and H. White. Oxford: Oxford University Press.
Teräsvirta, T. 1998. Modeling economic relationships with smooth transition regressions. In *Handbook of applied economic statistics*, ed. A. Ullah and D.E. Giles. New York: Dekker.
Teräsvirta, T. 2004. Smooth transition regression modeling. In *Applied time series econometrics*, ed. H. Lütkepohl and M. Krätzig. Cambridge: Cambridge University Press.
Teräsvirta, T. 2006a. Forecasting economic variables with nonlinear models. In *Handbook of economic forecasting*, vol. 1, ed. G. Elliott, C.W.J. Granger, and A. Timmermann. Amsterdam: North-Holland.
Teräsvirta, T. 2006b. Univariate nonlinear time series. In *Palgrave handbook of econometrics: volume 1, econometric theory*, ed. T.C. Mills and K. Patterson. Basingstoke: Palgrave Macmillan.
Teräsvirta, T., D. van Dijk, and M.C. Medeiros. 2005. Smooth transition autoregressions, neural networks, and linear models in forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting* 21: 755–774.
Tong, H. 1990. *Non-linear time series: A dynamical system approach*. Oxford: Oxford University Press.
Tsay, R.S. 1989. Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association* 84: 231–240.
Tsay, R.S. 1998. Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* 93: 1188–1202.
Tsay, R.S. 2002. *Analysis of financial time series*. New York: Wiley.
van Dijk, D., T. Teräsvirta, and P.H. Franses. 2002. Smooth transition autoregressive models – A survey of recent developments. *Econometric Reviews* 21: 1–47.
White, H. 2006. Approximate nonlinear forecasting methods. In *Handbook of economic forecasting*, ed. G. Elliott, C.W.J. Granger, and A. Timmermann. Amsterdam: North-Holland.

Thünen, Johann Heinrich von (1783–1850)

Jürg Niehans

JEL Classifications
B31

Thünen was born in Canarienhausen (Oldenburg) on 24 June 1783. He died on his estate Tellow (Mecklenburg), near Rostock, on 22 September 1850. His paternal ancestors were farmers; despite the 'von', they did not belong to the aristocracy.

After his father's early death, Thünen's mother married a timber merchant. The boy grew up in a small town on the northern seaboard, where he obtained a good, but short, high-school education. As an apprentice he got to know hard manual labour on a farm. There followed academic studies on all aspects of agronomy, including natural sciences, mathematics and economics, at the agricultural colleges of Gross-Flottbeck and Celle (where he heard Thaer) and at the University of Göttingen (where he read Adam Smith). Nevertheless, Thünen remained essentially a scientifically gifted autodidact. During this period (around 1803), he seems to have conceived the idea of his 'isolated state'.

Newly married to the daughter of a respected landowner, Thünen first operated a rented estate. In 1809, with the inheritance from his father, he bought from his brother-in-law the rather rundown estate of Tellow with about 1,200 acres of land. Though his heart was in his intellectual pursuits rather than in practical farming, he succeeded in gradually paying off his initial debt and in raising the value of his property, leaving to his four children a prosperous estate with ample liquid funds.

Like Quesnay, who came from a similar background, Thünen made the farm his economic paradigm. With the Physiocrats and Thaer, he belongs to those representatives of the Enlightenment who regarded improvements in

agriculture as the key to economic progress. For his estate he kept meticulous accounts, which he used to compute optimal solutions to management problems. He was a model employer with philanthropic, if somewhat paternalistic, ideas on social policy, who established a profit-sharing plan for his employees.

Of Thünen's *magnum opus,* 'The Isolated State with Respect to Agriculture and Political Economy', the first part, including the analysis of rent, location and resource allocation, appeared in 1826 after more than 20 years of work. The second part, containing the marginal productivity theory of distribution, only appeared in 1850. Additional papers, including important contributions on forestry, were published in 1863 by Thünen's biographer, H. Schumacher. All of this material is united in the third edition of 1875, but Waentig's later edition and also the English translations are limited to Part I and the first (and more important) half of Part II. Additional material was published by raeuer in his volume of selected works, which also includes a bibliography of Thünen's writings. The literary remains, including unpublished manuscripts, are preserved in the Thünen-Archiv at the University of Rostock.

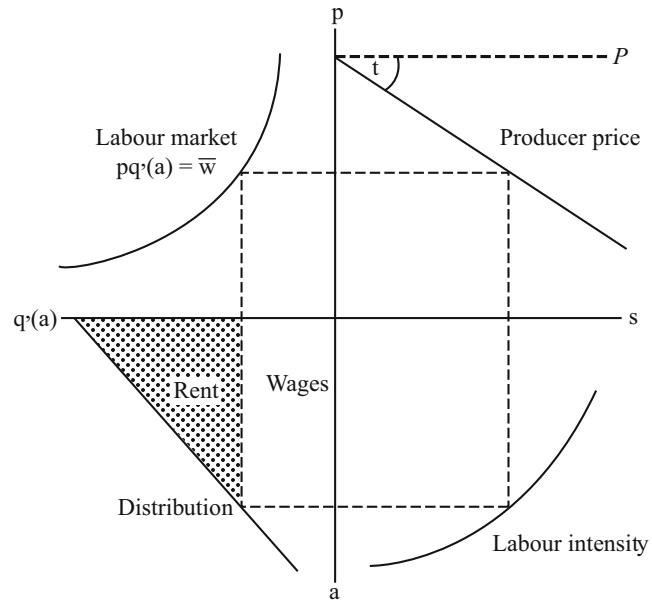
It has been said that Thünen was a prophet with little honour in any country, and even less in his own. This is inaccurate. It is true that he was at first disappointed about the reception of his book. Nevertheless, by 1827 he was an internationally known authority on agriculture, and the first edition was sold out within seven years. Tellow became a mecca for agronomists, attracting visitors from all over Europe. In 1830, Thünen was made a *doctor philosophiae honoris causa* by the University of Rostock. Politically a progressive liberal, he was elected to the National Assembly in Frankfurt in 1848, but could not attend because of his declining health. In the same year, the town of Teterow, with flags flying and bands playing, made him an honorary burgher. Like Quesnay, he died revered as a sage.

Thünen's scientific achievements are at different levels. In agronomy he made important contributions to the 'statics' of the soil, which are concerned with the steady state where fertility, by suitable crop rotation and fertilization, is maintained at an optimal level. In economics his most fundamental contribution is the method of deriving economic propositions from explicit optimizing models. By 1824 (as raeuer reports) this had led him to the differential calculus, which he may thus have been the first to apply to economic problems. At a time when German economists liked to criticize Adam Smith for his 'rationalism', Thünen criticized him for his lack of an explicit theory, which he undertook to provide. In mathematical elegance his contribution falls far short of Cournot's, but it exceeds the latter in breadth and depth. It makes Thünen one of the patron saints of modern economics.

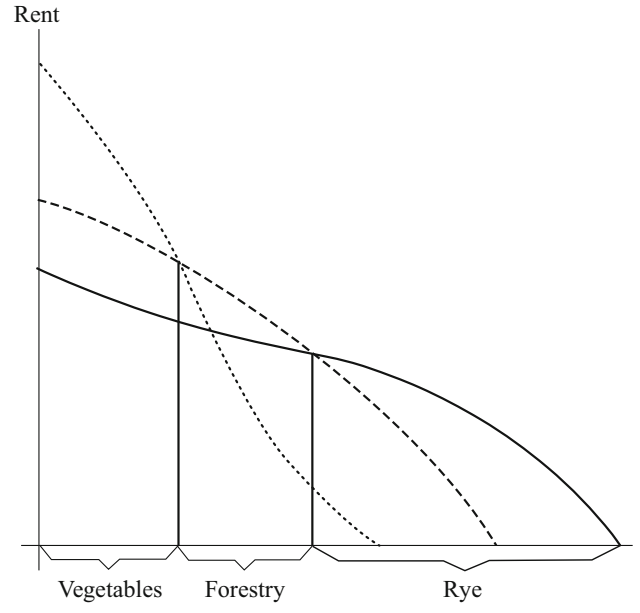
A more specific contribution is Thünen's theory of rent, location and resource allocation. In terms of modern economics, its elements can be summarized as follows. Suppose rye is sold in a central city at a given market price P. Production takes place on an unlimited plain of uniform fertility. Transportation to the market over s miles costs ts per bushel. The producer price, p, therefore, declines with increasing distance according to $p = P - ts$, as illustrated in the NE quadrant of Fig. 1. Output per acre depends on labour per acre according to $q = q(a)$. At each producer price, the manager selects his method of production in such a way that the marginal product of labour, $q'(a)$, evaluated at the producer price, equals the given (and uniform) wage rate, as described in the NW quadrant. The curve in the SW quadrant expresses the decline in the marginal product of labour as increasingly more labour-intensive methods are applied. The area 'below' the marginal product curve is total product. While the rectangle $q'(a)$ goes to wages, the shaded residual represents land rent. The curve in the SE quadrant, finally, shows labour intensity as a diminishing function of distance.

The solid curve in Fig. 2 graphs land rent from rye production as a diminishing function of distance. Similar rent–distance curves can be constructed for other products like vegetables or lumber. They are represented by, respectively, the dotted and the broken curve. At each distance the

Thünen, Johann Heinrich von (1783–1850), Fig. 1



Thünen, Johann Heinrich von (1783–1850), Fig. 2



farmer will plant the product promising the highest rent. This results in Thünen's famous rings. For a given product there may also be rings of different technologies.

In analysing the comparative statics of the 'isolated state', Thünen shows that lower transportation costs and more rapidly diminishing returns tend to increase the distance from the city at which

a good is produced or a technology is used. It is important to note that Thünen provides not only a theory of location, but also of factor intensities. That the relative efficiency of different technologies depends on market conditions is one of the main propositions he wanted to demonstrate.

The basic model is extended by Thünen in numerous directions. If the required quantities are given, the model determines their market prices. Since the rural workers do not generally pay the given city prices, their money wages will not actually be uniform. Freight costs may not be proportionate to distance. Substitutes and joint products are discussed. To the flows of agricultural products to the market centre, Thünen adds the reverse flows of consumer goods and means of production (like manure) and he pays attention to the unequal quality of the soil. The problem of the spatial distribution of several cities is raised, though not solved. It is finally shown that agricultural protection, by reducing the efficiency of land use, makes both parties worse off and that land taxes do not distort allocation. Despite its richness, Thünen's analysis remains partial in the sense that it does not determine a general spatial equilibrium. His notions about the price mechanism are crude. The long-winded discourse is replete with empirical calculations, relating Thünen's analysis to his account books down to the most minute details.

By applying his optimizing approach to factor inputs, Thünen became one of the originators of the marginal productivity theory of distribution. Using the Ricardian subterfuge of a rentless margin of cultivation, he explains his basic idea in the following words:

> Output p is the joint product of labour and capital. How should the share of each factor in the joint product be measured? We measured the effectiveness of capital by the increment in the output per worker due to an increase in the capital he works with. In this context, labour is constant, but capital is a variable magnitude. Suppose now that this procedure is continued, but in the reverse sense of considering capital as constant and labour as growing. In this case, in a large-scale operation, the effectiveness of labour (the contribution of the worker to output) is recognized from the increment in total output due to the augmentation of workers by one (II, §19).

As in Turgot, the output increments, from both capital and labour, are postulated to decline with an increasing factor input. The profit-maximizing entrepreneur will determine each factor input in such a way that the sales proceeds from the last unit are equal to the given factor price. This implies that at the point of minimum cost the ratio of factor prices is equal to the ratio of what today would be called their marginal products. The word 'marginal' does not occur, but the expressions 'margin' or 'limit' are constantly used.

From the laws that govern actual distribution, Thünen, deeply concerned with the 'social problem', proceeded to the laws that ought to govern it. This led him to the most controversial of his achievements, his famous 'natural wage' formula. In Thünen's economy, per capita output, p (measured in rye), depends on capital per worker, q (measured in terms of the tools a worker can make in a year). Output is divided between the wage, w, and the rental on capital, r, according to $p(q) = w + rq$. In sharp contrast to later notions, savings are supposed to come out of wages while property income is consumed. Specifically, savings are the excess of wages over some subsistence minimum, a. The economy is growing by the construction of new farms at the rent-less margin of cultivation.

With interest rate $(p - w)/wq$, the return on savings is

$$R = \frac{p - w}{wq}(w - a).$$

Thunen's 'natural wage' maximizes R on the assumption of fixed q (and thus p). By equating dR/dw to zero, the natural wage is easily determined as the geometric mean of p and a, $w = \sqrt{pa}$.

Thünen's cumbersome exposition has given rise to many misunderstandings. Some (including Marshall) argued that the correct interest rate would have been $(p - w)/q$. In this case, the natural wage turns out to be the arithmetic mean $w = \frac{1}{2}(p + a)$ (as already suggested by Knapp). This criticism would be valid for a one-sector economy in which q is simply a stock of

rye. Actually Thünen (as noted by Samuelson) considers a (rudimentary) two-sector economy in which capital goods are produced by labour only (at constant cost). Thünen is right, therefore, in valuing q at the wage rate w.

Another objection (raised, among others, by Wicksell and strongly reiterated by Samuelson) concerns the postulated constancy of q. After all, an increase in w presumably leads to an increase in q (and thus in p). Thünen anticipated this objection, for he supplemented his mathematical derivation, both verbally and by numerical examples, with a cogent explanation of how the overall maximum of R is to be found by searching over different q (and thus p). In fact, if output and marginal productivity wages are allowed to adjust to changes in q, the necessary condition for a maximum, as Dorfman (1986) showed, is again Thünen's square-root formula.

Many have thought Thünen's natural wage to be inconsistent with his own marginal productivity theory. If wages correspond to the marginal product of labour, how can they at the same time be expected to conform to some particular social ideal? This objection, however, loses its force once it is realized that Thünen (as observed by Dickinson) determined the capital/labour ratio at which the marginal productivity wage happens to be equal to his natural wage.

The fundamental objection to the natural wage formula is that it makes no sense for workers to be interested in the returns on their savings only. What Thünen seems to have been groping for, more than a century before Phelps et al., was a 'Golden Rule' of capital accumulation leading to some sort of optimal growth path (In a one-sector model, the arithmetic variant of the natural wage has indeed such properties; they are analysed in Samuelson 1986). He never got it right; in such an optimization problem, the savings parameter, a, can hardly be treated as given. Thünen regarded his formula as important enough to have it engraved on his tombstone in the churchyard of elitz. It commemorates a brilliant failure.

Part III of the 'Isolated State' is concerned with the efficiency of forest management, thereby extending the incomplete treatment in Part I. The detailed analysis of the optimal spacing of trees is of interest mainly to forest engineers. In analysing the optimal rotation period, however, Thünen makes another important contribution to economic theory. He had already pointed out in Part I that the value of a forest should not be measured by the sales value of the timber if the trees are cut today, but rather by the present value of the timber if the trees are cut and sold at the end of the optimal rotation period. In an efficient operation, the latter exceeds the former; if not, the trees should be cut at once. Efficient forest management is thus interpreted as a problem of capital and interest, providing economic theory with one of its most fruitful paradigms.

Thünen's optimality criterion, in contrast to Wicksell and Fisher, is not the equality of the marginal product of capital and the rate of interest, which, by disregarding the value of land, results in cutting trees too late. As Manz (1986) has shown, Thünen was probably the first to use the correct criterion of maximal land rent, which shortly afterwards was so brilliantly developed by Faustmann. The formula derived in Part III is flawed by incorrect discounting, and the exposition is clumsy. Nevertheless, with respect to substantive content, the capital theory implied in Thünen's forest model is superior to öhm-awerk's and it was not surpassed in economic science before Wicksell.

See Also

- ▶ Location Theory
- ▶ Marginal Productivity Theory
- ▶ Monocentric Versus Polycentric Models in Urban Economics

Selected Works

1826–63. Der isolierte Staat in eziehung auf Landwirtschaft und Nationalökonomie. Pt I:

Untersuchungen über den Einfluss, den die Getreidepreise, der Reichtum des odens und die Abgaben auf den Ackerbau ausüben. Hamburg: Perthes, 1826. 2nd ed., Rostock: Leopold, 1842. Pt II: Der naturgemässe Arbeitslohn und dessen Verhältniss zum Zinsfuss und zur Landrente. Rostock: Leopold; 1st section 1850; 2nd section 1863. Pt III: Grundsätze zur estimmung der odenrente, der vorteilhaftesten Umtriebszeit und des Werts der Holzbestände von verschiedenem Alter für Kieferwaldungen, Rostock: Leopold, 1863.

1951. *Ausgewählte Texte*, ed. Walter raeuer, Die grossen Sozialökonomen, vol. VII. Meisenheim: Hain, 1951. English translations: Pt I: *Von Thünen's isolated state,* ed. Peter Hall, trans. Carla M. Wartenberg. Oxford: Pergamon Press, 1966. Pt II, section 1, in ernard W. Dempsey, *The frontier wage*. Chicago: Loyola University Press, 1960.

Bibliography

Braeuer, W. 1950. Der Mathematiker-Oekonom. Zur Erinnerung an Johann Heinrich von Thünen. *Kyklos* 4 (2/3): 150–171.

Buhr, W. 1983. Mikroökonomische Modelle der von Thünenschen Standorttheorie. *Zeitschrift für Wirtschafts- und Sozialwissenschaften* 103 (3): 587–627.

Bulow, F. 1958. Johann Heinrich von Thülow, F. 1958. Johann Heinrich von Thünen als forstwirtschaftlicher Denker. Zur Erinnerung an den 175. Geburtstag Johann Heinrich von Thünens am 24. Juni 1783. *Weltwirtschaftliches Archiv* 80 (2): 183–233.

Dickinson, H.D. 1969. Von Thünen's economics. *Economic Journal* 79 (316): 894–902.

Dorfman, R. 1986. Comment: P.A. Samuelson, 'Thünen at two hundred'. *Journal of Economic Literature* 24.

Ehrenberg, R. 1905a. Johann Heinrich von Thünen. *Thünen-Archiv. Organ für exakte Wirtschaftsforschung* 1.

Ehrenberg, R. 1905b. Thünen's erste wirtschaftswissenschaftliche Studien. *Thünen-Archiv. Organ für exakte Wirtschaftsforschung* 1.

Ehrenberg, R. 1905c. Thünen und Thaer. *Thünen-Archiv Organ für exakte Wirtschaftsforschung* 1.

Ehrenberg, R. 1909. Entstehung und Wesen der wissenschaftlichen Methode Johann Heinrich von Thünen's. *Thünen-Archiv. Organ für exakte Wirtschaftsforschung* 2.

Franz, G., ed. 1958. Johann Heinrich von Thünen. *Zeitschrift für Agrargeschichte und Agrarsoziologie* (Special issue) 6. Frankfurt: DLG.

Knapp, G.F. 1865. *Zur Prüfung der Untersuchungen Thünen's über Lohn und Zinsfuss im isolierten Staate*. Braunschweig: Vieweg.

Krzymowski, R. 1928. Graphical presentation of Thünen's theory of intensity. *Journal of Farm Economics* 10: 461–482.

Leigh, A.H. 1946. Von Thünen's theory of distribution and the advent of marginal analysis. *Journal of Political Economy* 54: 481–502.

Manz, P. 1986. Forestry economics in the steady-state. The contribution of JH von Thünen. *History of Political Economy* 18 (2): 281–290.

Moore, H.L. 1895. Von Thünen's theory of natural wages. *Quarterly Journal of Economics* 9(April): 291–304; (July): 388–408.

Salin, E. 1926. Der isolierte Staat 1826–1926. *Zeitschrift für die gesamte Staatswissenschaft* 81 (3): 410–431.

Samuelson, P.A. 1983. Thünen at two hundred. *Journal of Economic Literature* 21 (4): 1468–1488.

Samuelson, P.A. 1986. Yes to Robert Dorfman's vindication of Thünen's natural-wage derivation. *Journal of Economic Literature* 24.

Schneider, E. 1959. Johann Heinrich von Thünen und die Wirtschaftstheorie der Gegenwart. *Schriften des Vereins für Sozialpolitik* 14: 14–28.

Schumacher, H. 1868. *Johann Heinrich von Thünen. Ein Forscherleben*. Rostock: Leopold.

Seedorf, W., and H.-J. Seraphim, eds. 1933. *Johann Heinrich von Thünen zum 150. Geburtstag. Versuch der Würdigung einer Forscherpersönlichkeit*. Rostock: Hinstorffs.

Woermann, E. 1959. Johann Heinrich von Thünen und die landwirtschaftliche etriebslehre der Gegenwart. *Schriften des Vereins für Sozialpolitik* 34: 28–45.

Tiebout Hypothesis

Dennis Epple

Abstract

Tiebout (J Polit Econ 64: 416–424, 1956) argued that efficient local public good provision would emerge as households choose among communities offering different local public goods bundles. This article highlights research linking the Tiebout hypothesis to real-world local political jurisdictions, the first such

link having been forged by the pioneering contribution of Oates (J Polit Econ 77: 957–971, 1969). Subsequent research has studied voting over tax and expenditure policies within municipalities in a metropolitan area, and sorting of the metropolitan population both within and across those municipalities. This research has provided the foundation for econometric analysis and policy applications.

Keywords

Clubs; Collective choice; Demography; Equity vs efficiency; Exit and voice; Fiscal zoning; Income stratification; Jurisdictional competition; Land use planning; Local government; Local public goods; Multi-community equilibrium; Neighbourhood effects; Peer effects; School choice; School vouchers; Tiebout hypothesis; Tiebout, C.

JEL Classifications

R51

In his famous paper, Charles Tiebout (1956) argued that there were realistic conditions under which local public goods would be provided efficiently; an efficient allocation would emerge as each household selected the community providing the public good levels most closely aligned to its preferences. This has come to be known as the 'Tiebout hypothesis' and the related Tiebout community-choice mechanism has been dubbed 'voting with your feet'. This article focuses on research linking the Tiebout hypothesis to local political jurisdictions.

Oates (1969) gave Tiebout's hypothesis empirical content. He reasoned that if households selected among communities in the way Tiebout conjectured, 'capitalization' should result. That is, *ceteribus paribus*, housing prices should be higher in communities with high levels of public good provision and lower in communities with high tax rates. Oates tested and found support for these predictions using data for municipalities in New Jersey. His paper led to an explosion of research on capitalization and launched research into a variety of related aspects of the Tiebout hypothesis.

Choice Within Jurisdictions

Tiebout was largely silent about how communities would settle on their levels of public good provision, though he emphasized parallels with market provision. Study of this market-based approach is largely the domain of the theory of clubs. Another approach, initiated by Barr and Davis (1966), focuses on collective choice within communities. In the terms of Albert Hirschman (1970), Tiebout emphasized 'exit' while Barr and Davis emphasized 'voice'. Much research followed. Bergstrom and Goodman (1973) formalized estimation of demand for local public goods. Romer and Rosenthal (1979) investigated the role of agenda setters. Micro-level estimation of demand for local public goods was undertaken by Bergstrom et al. (1982).

Goldstein and Pauly (1981) observed that neglect of self-selection of households into communities would potentially bias estimates of demands for local public goods. Work followed linking intra-community choice and inter-community choice, beginning with Rubinfeld et al. (1987), and continuing with research on models of multi-community equilibrium.

Equilibrium Among Jurisdictions

Not surprisingly, households with higher incomes tend to prefer communities with high levels of local public goods. It is natural to ask whether income stratification can be sustained in equilibrium. Building on the work of Ellickson (1971), Westhoff (1977) proves existence of equilibrium in a model with sorting across communities and voting within communities. Westhoff's model is extended to incorporate housing markets by Epple et al. (1984), who demonstrate that income-stratified equilibria can be sustained by differentials across communities in the price per unit housing of services. Fernandez and Rogerson (1998) add an important dynamic feature, with community education spending by each generation affecting incomes of the succeeding generation.

While households tend to sort by income across communities, there is much income variation within

jurisdictions, even within small neighbourhoods (Hardman and Ioannides 2004). Several approaches seek to capture this intra-community heterogeneity as an outcome in multi-community equilibrium. One approach emphasizes heterogeneity in preferences as well as incomes. Structural estimation of multi-community equilibrium models embodying such heterogeneity is undertaken by Epple and Sieg (1999) and Epple et al. (2001). This framework is applied to study large-scale policy change by Sieg et al. (2004).

An alternative approach emphasizing heterogeneity and durability of housing is developed by Nechyba (1997). Nechyba has extended and applied this framework to study important policy issues, with particular emphasis on school choice and vouchers (Nechyba 2000). Structural estimation taking Nechyba's model as a point of departure is undertaken by Ferreyra (2005) who extends the model to include heterogeneity in household tastes, including tastes for sectarian and non-sectarian schools.

Still another approach, by Bayer et al. (2004), permits detailed investigation of the way a household's own demographic characteristics affect preferences with respect to the demographic composition of communities and the quality of local public goods. Bayer et al. (2005) apply this framework to estimate household preferences for community composition and to investigate the extent to which sorting within a metropolitan area is driven by preferences for education.

Residents of a jurisdiction may affect the public goods provided therein via 'neighbourhood effects' and 'peer effects'. Peer effects are introduced into a multijurisdictional model by deBartolome (1990). While research on the Tiebout model emphasizes choice among municipalities in a metropolitan area, there is also population sorting within municipalities, especially central cities. Benabou (1996) and Durlauf (1996) study how peer effects influence such sorting, and the economic consequences of such sorting. Multi-community models increasingly emphasize peer effects (Nechyba 2000; Epple and Romano 2003; Bayer et al. 2004; Rothstein 2006; Sethi and Somanathan 2004; Ferreyra 2005).

Equity and Efficiency

Local governments impose many restrictions on land use (Fischel 1985). Hamilton (1975) emphasizes the potential efficiency-enhancing role of zoning. Other research investigates 'fiscal zoning', the allegation that jurisdictions use zoning to restrict entry by households who would contribute less in taxes than the cost of public services they would consume. Early contributions are in Mills and Oates (1975). Multi-community models with community residents choosing zoning by majority rule are developed by Fernandez and Rogerson (1997) and Calabrese et al. (2005). Computational results in the latter reveal that zoning can enhance efficiency, but also support critics who argue that fiscal zoning benefits wealthy households at the expense of poorer households. Henderson (1985) emphasizes the role of the private sector in community development. Henderson and Thisse (2001) focus on the role of developers in determining the character of the housing stock in communities. Glaeser and Gyourko (2002) conclude that land use restrictions play a major role in driving up housing prices in some areas of the United States, particularly California and some eastern cities.

The essence of the Tiebout hypothesis is that localities provide differing public good bundles to reflect variation within the population in tastes and incomes. Education is arguably the most important locally provided good, and efficiency arguments favouring decentralized provision in Tiebout equilibrium lie in uneasy juxtaposition with equity arguments favouring more centralized provision to increase equality of educational opportunity. US Courts have mandated intervention in many states to achieve greater equality of spending (Evans et al. 1998).There is growing emphasis in research (Duncombe and Yinger 1998) and the courts on policies designed to yield greater equality of educational outcomes. There is also increasing emphasis on incentive systems that might stimulate efficient provision of public education (Ladd 1996) and increasing recognition that interventions intended to achieve greater equity may affect both political support for public education and effectiveness of provision.

Oates (2006, p. 42) puts the matter succinctly: 'There seems to be an inevitable tension here.' Policies promoting equity need to be designed to harness local incentives for effective provision while also recognizing their impact on political support for public provision.

See Also

- Clubs
- Educational Finance
- Local Public Finance
- Social Interactions (Empirics)
- Social Interactions (Theory)
- Urban Economics
- Urban Political Economy

Acknowledgment In writing this article, I have benefited greatly from Oates (2006) and Fischel (2006). More extensive treatment is provided in Ross and Yinger (1998), Scotchmer (2002) and Epple and Nechyba (2004).

Bibliography

Barr, J., and O. Davis. 1966. An elementary political and economic theory of expenditures of state and local governments. *Southern Economic Journal* 33: 149–165.

Bayer, P., R. McMillan, and K. Rueben. 2004. *An equilibrium model of sorting in an urban housing market*. Working paper no. 10865. Cambridge, MA: NBER.

Bayer, P., F. Ferreira, and R. McMillan. 2005. *Tiebout sorting, social multipliers and the demand for school quality*. Working paper. Yale University.

Benabou, R. 1996. Equity and efficiency in human capital investment: The local connection. *Review of Economic Studies* 63: 237–264.

Bergstrom, T., and R. Goodman. 1973. Private demands for public goods. *American Economic Review* 63: 280–296.

Bergstrom, T., D. Rubinfield, and P. Shapiro. 1982. Micro-based estimates of demand functions for local school expenditures. *Econometrica* 50: 1183–1205.

Calabrese, S., D. Epple, and R. Romano. 2005. *On the political economy of zoning*. Working paper. Carnegie Mellon University.

deBartolome, C. 1990. Equilibrium and inefficiency in a community model with peer group effects. *Journal of Political Economy* 98: 110–133.

Duncombe, W., and J. Yinger. 1998. School finance reform: Aid formulas and equity objectives. *National Tax Journal* 51: 239–262.

Durlauf, S. 1996. A theory of persistent income inequality. *Journal of Economic Growth* 1: 75–93.

Ellickson, B. 1971. Jurisdictional fragmentation and residential choice. *American Economic Review Papers and Proceedings* 61: 334–339.

Epple, D., and T. Nechyba. 2004. Fiscal decentralization. In *Handbook of regional and urban economics*, ed. J. Henderson and J.-F. Thisse, vol. 4. Amsterdam: North-Holland.

Epple, D., and R. Romano. 2003. Neighborhood schools, choice, and the distribution of educational benefits. In *The economics of school choice*, ed. C. Hoxby. Cambridge, MA: NBER.

Epple, D., and H. Sieg. 1999. Estimating equilibrium models of local jurisdictions. *Journal of Political Economy* 107: 645–681.

Epple, D., R. Filimon, and T. Romer. 1984. Equilibrium among local jurisdictions: Toward an integrated treatment of voting and residential choice. *Journal of Public Economics* 24: 281–308.

Epple, D., T. Romer, and H. Sieg. 2001. Interjurisdictional sorting and majority rule: An empirical analysis. *Econometrica* 69: 1437–1465.

Evans, W., S. Murray, and R. Schwab. 1998. Education finance reform and the distribution of education resources. *American Economic Review* 88: 789–812.

Fernandez, R., and R. Rogerson. 1997. Keeping people out: Income distribution, zoning, and the quality of public education. *International Economic Review* 38: 23–42.

Fernandez, R., and R. Rogerson. 1998. Public education and income distribution: A dynamic quantitative evaluation of education-finance reform. *American Economic Review* 88: 813–833.

Ferreyra, M. 2005. *Estimating the effects of private school vouchers in multi-district economies*, Working paper. Carnegie Mellon University.

Fischel, W. 1985. *The economics of zoning laws: A property rights approach to American land use controls*. Baltimore: Johns Hopkins University Press.

Fischel, W. 2006. Footloose at fifty: An introduction to the Tiebout anniversary essays. In *The Tiebout Model at fifty: Essays in public economics in honor of Wallace Oates*, ed. W. Fischel. Cambridge, MA: Lincoln Institute of Land Policy.

Glaeser, E., and J. Gyourko. 2002. *The impact of zoning on housing affordability*, NBER working paper no. 8835. Cambridge, MA: NBER.

Goldstein, G., and M. Pauly. 1981. Tiebout bias on the demand for local public goods. *Journal of Public Economics* 16: 131–143.

Hamilton, B. 1975. Zoning and property taxation in a system of local governments. *Urban Studies* 12: 205–211.

Hardman, A., and Y. Ioannides. 2004. Neighbors' income distribution: Economic segregation and mixing in US urban neighborhoods. *Journal of Housing Economics* 13: 368–382.

Henderson, J. 1985. The Tiebout model: Bring back the entrepreneurs. *Journal of Political Economy* 93: 248–264.
Henderson, J., and J.-F. Thisse. 2001. On strategic community development. *Journal of Political Economy* 109: 546–569.
Hirschman, A. 1970. *Exit, voice, and loyalty*. Cambridge, MA: Harvard University Press.
Ladd, H. 1996. *Holding schools accountable: Performance-based reform in education*. Washington, DC: Brookings Institution.
Mills, E., and W. Oates, eds. 1975. *Fiscal zoning and land use controls*. Lexington: Heath-Lexington Books.
Nechyba, T. 1997. Existence of equilibrium and stratification in local and hierarchical public good economies with property taxes and voting. *Economic Theory* 10: 277–304.
Nechyba, T. 2000. Mobility, targeting and private school vouchers. *American Economic Review* 90: 130–146.
Oates, W. 1969. The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy* 77: 957–971.
Oates, W. 2006. The many faces of Tiebout. In *The Tiebout Model at fifty: Essays in public economics in honor of Wallace Oates*, ed. W. Fischel. Cambridge, MA: Lincoln Institute of Land Policy.
Romer, T., and H. Rosenthal. 1979. Bureaucrats versus voters: On the political economy of resource allocation by direct democracy. *Quarterly Journal of Economics* 93: 563–587.
Ross, S., and J. Yinger. 1998. Sorting and voting: A review of the literature on urban public finance. In *Handbook of regional and urban economics*, ed. P. Cheshire and E. Mills, vol. 3. Amsterdam: North-Holland.
Rothstein, J. 2006. Good principals or good peers? Parental valuation of school characteristics, Tiebout equilibrium, and the incentive effects of competition among juridisctions. *American Economic Review* 96: 1333–1350.
Rubinfeld, D., P. Shapiro, and J. Roberts. 1987. Tiebout bias and the demand for local public schooling. *Review of Economics and Statistics* 69: 426–437.
Scotchmer, S. 2002. Local public goods and clubs. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein, vol. 4. Amsterdam: North-Holland.
Sethi, R., and R. Somanathan. 2004. Inequality and segregation. *Journal of Political Economy* 112: 1296–1321.
Sieg, H., V. Smith, H. Banzhaf, and R. Walsh. 2004. Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *International Economic Review* 45: 1047–1077.
Tiebout, C. 1956. A pure theory of local public expenditures. *Journal of Political Economy* 64: 416–424.
Westhoff, F. 1977. Existence of equilibrium in economies with a local public good. *Journal of Economic Theory* 14: 84–112.

Tiebout, Charles Mills (1924–1968)

Murray Milgate

Keywords

Foreign-trade multiplier; Local public finance; Public goods; Regional development; Tiebout, C. M.; Urban economics

JEL Classifications

B31

Tiebout was born in Norwalk, Connecticut, took his BA from Wesleyan University in 1950, and his Ph.D. from the University of Michigan in 1957. After holding appointments at Northwestern (1954–8) and the University of California at Los Angeles (1958–62), he became professor of economics and business administration at the University of Washington at Seattle in 1962. He died on 16 January 1968. By far his most important work was his 'Pure Theory of Local Public Expenditure', which appeared in the October number of the *Journal of Political Economy* for 1956, from which is derived the so-called Tiebout hypothesis, which is the subject of a separate article in this Dictionary. However, his work on problems in regional and urban economics was more extensive than this one theoretical article on the local provision of public goods might suggest.

For example, in the volume of the *Journal of Political Economy* that carried his essay on public goods, there also appeared a paper which examined the effects of export growth on the pattern of regional economic development. This analysis represents an attempt to apply a Keynesian model of income determination to regional development. Tiebout argued that exports are only one of a number of sources that act to determine the growth of regional income, and through what appears to be the first application of the foreign-trade multiplier to regional analysis he attempts to reach conclusions as to the relative significance of

regional exports vis-à-vis regional demand as a source of income generation. Principal among these is that export-led regional growth is likely to be most effective when the regional base is small.

This paper was followed by work on the construction and use of regional and inter-regional input-output models (1957), which itself produced empirical investigations into the regional distribution of economic activity in the American states of California (1963) and of Washington (1969). The results of these studies were still appearing after Tiebout's early death – especially to be noted in this regard is his inter-regional input-output model of the linkages between the economies of Washington and California (1970). Add to this his work on the regional impact of the federal government's dispensation of its defence and space budgets (1964), and it becomes fairly clear that to record Tiebout's name solely in regard to his contribution to the pure theory of public goods would be to present a rather one-sided picture of his scientific interests.

Selected Works

1956a. A pure theory of local public expenditures. *Journal of Political Economy* 64: 416–424.

1956b. Exports and regional economic growth. *Journal of Political Economy* 64: 160–164.

1957. Regional and inter-regional input–output models: An appraisal. *Southern Economic Journal* 24: 140–147.

1960a. Community income multipliers: A population growth model. *Journal of Regional Science* 2: 75–84.

1960b. Economies of scale and metropolitan governments. *Review of Economics and Statistics* 42: 442–444.

1963. (With W.L. Hansen.) An intersectoral flows analysis of the Californian economy. *Review of Economics and Statistics* 45: 409–418.

1964. (With R.S. Peterson.) Measuring the impact of regional defence-space expenditures. *Review of Economics and Statistics* 46: 421–428.

1969. An empirical regional input-output model: The state of Washington. *Review of Economics and Statistics* 51: 334–340.

1970. Inter-regional input-output: An empirical California–Washington model. *Journal of Regional Science* 10: 133–152.

Tight Money

A. B. Cramp

This term, like many others in financial economics, has no single and precise meaning because over the years debate has come inevitably to reflect the complexity of financial realities. But some pointers towards definition may usefully be made. First, the reference is to tightness of the supply of money relative to the demand for it. Recognition of this point, elementary yet easily overlooked in the heat of practical controversy, immediately indicates the problem that tightness/ease is typically difficult to identify – much less to measure – with any confidence. For example, while interest-rate levels are frequently regarded as appropriate indicators, low and/or falling rates can indicate monetary ease if they reflect supply expanding more rapidly than demand, but could reflect tightness if demand has contracted – say because of collapsing sales and profits – so that equilibrium interest rate levels would be lower still.

Second, we must ask *what* is in tight supply. In some earlier phases of discussion, notably in the late 19th and early 20th centuries, attention was focused on the inelasticity of the supply of currency under gold-standard régimes, in the face of rising demand to finance transactions as boom conditions developed, and of the need to bolster confidence as those conditions gave way to liquidity difficulties around the peak of the cycle. Thus (over-)tight money was seen as evidence of failure to solve one aspect of the problem of sound money.

The abandonment of the gold standard, and the development of smoothing techniques by central banks, caused this particular difficulty to fade. In later phases, attention has shifted towards inelasticity of the supply of bank credit, whether in the sense of the flow of bank lending or in that of the stock of bank deposits, and whether occurring as a result of the autonomous operation of the market or of policy measures taken by the monetary authorities. Though definition is often not made explicit, most discussion probably has made implicit reference to policy-induced restriction of the flow of bank lending, from the supply side. Such restriction would normally be accompanied by a tendency for short-term interest rates to rise relative to what would otherwise have occurred, but probably also by expansion of Keynes's 'fringe of unsatisfied borrowers' in imperfect credit markets. Dear money, that is to say, may be caused by tightness of money, but tightness is not necessarily fully reflected in price.

This last point draws our attention to the undoubted fact that the impact of tight money is by no means likely to be evenly distributed over the population of transactors in an economy. Awareness of this fact has doubtless been one reason prompting the question (G.L. Bach, in Carson, 1963), 'How discriminatory is tight money?' Before outlining Bach's own answer to this question, it is necessary to emphasize that his discussion relates to the earlier of two phases into which the post-1945 history of this issue naturally divides. In that earlier phase, covering roughly the years to 1970, 'tightness' of money was both an occasional and (by later standards) a distinctly modest phenomenon. Since 1970, in the era of anti-inflationary monetary targets, tightness has tended to be both more continuous and more severe.

Bach's evidence in fact related primarily to the years 1955–7, and to the USA. He found little evidence to support one common hypothesis, that tight money caused discrimination against small business borrowers in favour of larger ones – a conclusion reinforced by indications that, in the conditions of this earlier phase at least, small businesses could respond to restrictions on the supply of bank credit by taking relatively more net trade credit. On the other hand, the investigation provided additional support for another form of discrimination that is so widely recognized, especially for the earlier phase, as to be hardly open to dispute: because interest rates in the mortgage market were, for institutional reasons, 'sticky' relative to market rates, the overall availability of mortgage finance was undoubtedly somewhat reduced in periods of relative monetary tightness.

One other kind of discriminatory effect of tight money and higher-than-otherwise interest rates has been widely discussed. This is the effect on income distribution. Banks, for example, have been supposed to gain from tight money, on account of the so-called endowment element arising from their zero-interest deposits on current account, so that generally rising interest rates tended to increase the differential between average income on assets and average cost of liabilities. This argument was doubtless valid, though its effects were offset in periods of monetary ease. Its importance has in any case declined with the spread (from the early 1980s) of mechanisms for paying interest on current accounts.

More generally, tight money has been argued to have differential implications for creditors and debtors. Their nature, however, is ambiguous. On the creditor side, for example, new lenders benefit from higher interest rates, but existing lenders may suffer from decline in the market value of fixed-interest assets. These changes are mirrored on the debtor side, but individuals are frequently both debtors and creditors, and groups of transactors (e.g. the company sector) will certainly include both. So, while it must be recognized that tight money (like the inflationary situation it is usually designed to counter) will cause some to gain and others to lose, no cogent generalization is possible.

In any case, the major controversy over tight money concerns not its differential, but its general, impact. Any discriminatory effects are more or less incidental to its normal purpose (when it results from policy decisions) of reducing the level of aggregate demand and associated

inflationary pressures in the economy. In examining this matter, it is again desirable to distinguish between the earlier and the later phases of the post-1945 era.

Before 1970, Gurley and Shaw (1960) had provided rigorous exposition of a theoretical standpoint, applied in the 1959 Radcliffe Report to then-contemporary UK institutions and conventions, which postulated that the effects of limited tightness of money could be circumvented by economic agents with relative ease. The main focus of the argument was on the operation of financial intermediaries other than banks, many of which (notably, but far from exclusively, UK building societies and US savings and loan associations) issued liabilities that were relatively close substitutes for bank deposits – certainly as stores of value, and to some extent also as exchange media. These non-bank financial intermediaries could attract additional deposits by relatively small interest-rate increases (the bank-deposit preference schedule was interest-elastic owing to the existence of close substitutes), and use the funds to finance loans to agents affected by a squeeze on bank credit. So, the effects of tight money would be neutralized by rerouting of fund flows, reflected in rising velocity of circulation, and in interest-rate increases that tended to be insignificant given a relatively interest-inelastic marginal efficiency of capital schedule.

The force of such arguments clearly depended, *inter alia*, on the maintenance of confidence by (real) investors, and that confidence in turn rested in part on the absence of threats to the stability of the financial system. Periods of more severe monetary tightness since about 1970 have been characterized by such threats (cf. the 1974 secondary banking crisis in the UK and elsewhere), in a manner giving practical force to theoretical ideas associated – in the modern period – primarily with the writings of Hyman Minsky, progenitor of the so-called 'financial instability thesis'. One central strand of these ideas relates to the pattern of company finance during the cycle upswing. It sees non-financial companies as prone, in this phase, to euphoric expectations which result in financing patterns (relatively more external finance compared with internal, bond finance compared with equity, short-term borrowing compared with long) causing deterioration of the financial strength of balance-sheet positions. Because rather similar influences affect the condition of financial intermediaries, bank and non-bank, the liquidity problems typical of the cycle peak can easily cause a finance shortfall resulting in inability to meet by due date cash-payment commitments that prove to have been based on over-optimistic expectations. In such 'fragile' circumstances, tight money can cause insolvency and bankruptcy, tending to spread because of interdependence between transactors (A's ability to fulfil commitments to B is undermined if C fails to fulfil commitments to A . . .). Therefore central banks will have powerful incentive to acquiesce in easy money conditions – which, if crisis is indeed thereby forestalled, may be followed by a new inflationary boom in a sequence that monetarist models inevitably misinterpret.

The argument just outlined explains why central banks tend, at least in the late upswing and peak phases of the cycle, to give priority to what the present writer has labelled *support* objectives as opposed to those *control* objectives whose potential has been inappropriately overemphasized by monetarist/neoclassical theorizing. In turn, the necessary concern with support objectives explains why, even in the later postwar phase of supposedly continuous tight contrainflationary money supply targets, the effective degree of tightness (in any case impossible to measure with any precision) has probably tended to be less than it might appear on a superficial view. The Bank of England in the first half of the 1980s, for example, has responded to high private sector demand for bank loans by 'overfunding' (selling more government debt than needed to finance current spending requirements); thus negative bank lending to the public sector, in the face of high levels of lending to the private sector, facilitates the achievement of money supply targets. But to offset the resultant squeeze on bank reserve ratios, the Bank has been ready to buy commercial bills of exchange from the banks, adding to its so-called 'bill mountain', but permitting private sector borrowing to continue with little effective check.

Emphasis on this kind of sophisticated combination of tightening/loosening measures, however, should not be exaggerated to the point of arguing that tightness of money has never been sufficient to frustrate private sector spending plans. There have been periods (during what we have called the later, post-1970 phase) in which, partly to persuade markets to absorb enough government debt to enable money-supply targets to be met, apparent real interest rates have been at levels extremely high by reference both to long-term historical experience and to current profit levels. This may well have been due as much to market reaction against earlier experience of negative real interest rates as to the direct effects of current monetary policy, so that it is debatable how far the phenomenon should be regarded as the outcome of tight money; it might more appropriately be regarded as the price of earlier weakness in such areas as fiscal and incomes policies. Be that as it may, there is little doubt that in such periods private-sector agents have suffered considerable damage (see, e.g. Miller and Lonie, 1978). High real interest rates reduce the profitability of the operations of existing borrowers. *Falling* inflation rates adversely affect cash-flow projections for operations proposed to be financed by new loans. There is a general tendency to decline of credit-worthiness, which explains how valid cries of pain from borrowers and valid claims by banks to be meeting all credit-worthy demands can co-exist, and probably also is part of the explanation of generally rising levels of company liquidations and bankruptcies. The fruits of inflation are bitter, but on the view taken here monetarist doctrine is inadequate to explicate either inflation's cause or its cure.

See Also

► Dear Money

Bibliography

Carson, D. (ed.). 1963. *Banking and monetary studies*. Homewood: Richard D. Irwin.

Gurley, J.G., and E.S. Shaw. 1960. *Money in a theory of finance*. Washington, DC: Brookings Institution.

Miller, E., and A. Lonie. 1978. *Micro-economic effects of monetary policy*. London: Martin Robertson.

Time Consistency of Monetary and Fiscal Policy

Paul Klein

Abstract

Why do even benevolent policymakers frequently break their promises? Kydland and Prescott (1977) discovered that, when outcomes depend on expectations, rational policy choices typically depend on whether (*a*) the policymaker takes into account the constraint that the expected policy is the actual policy or (*b*) she takes expectations as given. A government that *commits* itself to a policy takes this constraint into account, a government that acts at its discretion does not. Since the commitment policy leads to a better outcome, there is the temptation to announce it and then to abandon this policy. This is the time inconsistency problem.

Keywords

Asymmetric information; Central bank independence; Commitment; Expectations; Friedman rule; Government budget constraint; Inflation; Inflationary expectations; Markov perfect equilibrium; Phillips curve; Public debt; Ramsey equilibrium; Rational expectations; Reputation; Second best; Sticky prices; Sustainable equilibrium; Time consistency of monetary and fiscal policy; Time inconsistency

JEL Classifications

D4; D10

A (possibly time- and state-contingent) strategy is said to be time inconsistent if an agent finds it

optimal from the point of view of some initial period 0 but finds it suboptimal in some subsequent period *t*. Time inconsistency can obviously arise if the government has time-varying preferences because of alternations of government, as shown in Persson and Svensson (1989). However, as Kydland and Prescott (1977) discovered, the time-inconsistency problem is a pervasive feature of environments with a single benevolent policymaker taking decisions over time. This happens even though the policymaker has stable preferences and even in situations where there is no apparent conflict of interest – though the emphasis there should perhaps be on the word 'apparent'. This means that everyone in the economy can often be made better off if the policymaker gains access to a commitment technology – a mechanism that forces him or her to keep his or her promises.

As pointed out by Fischer (1980), what these environments have in common is (*a*) that the Pareto optimal allocation is not implementable, (*b*) that the behaviour of the private sector depends on its expectations about future government behaviour, and (*c*) that the government and a typical individual do not share the same preferences. The third of these sounds stronger than it is. It is consistent with the benevolence of the policymaker, that is, that she maximizes the utility of a representative individual. All that it requires is a minimal amount of selfishness on the part of a typical individual – for example, she doesn't internalize the government's budget constraint.

It is important to note that either (*a*) or (*b*) on its own is not problematic. If the Pareto optimum can be achieved and I declare a policy at the beginning of time that is consistent with this Pareto optimum, I have no reason to deviate in the future since nothing better is feasible. On the other hand, suppose the Pareto optimum cannot be achieved but current behaviour does not depend on future policy. Then the policymaker will make the right trade-off in each period and the second best will be achieved.

However, if both features are present at the same time, then the policy that achieves the second best will typically be time inconsistent: when choosing policy for period *t*, the policymaker faces different incentives in period 0 from the incentives faced in period *t*. In period 0 she rationally takes into account the effects on expectations; in period *t* it is no longer rational to do that, since expectations in the past are bygones. The temptation to bring the economy closer to the first-best renders the second-best solution time inconsistent, and rational expectations force the economy into a Pareto inferior third-best equilibrium.

The Phillips Curve and Inflation Bias

The central example in Kydland and Prescott (1977) is a central bank setting inflation in an environment with an expectations-augmented Phillips curve. This environment has the feature that inflation surprises lead to deviations of output from its 'natural' rate. The authors then assume that a positive inflation surprise is good, since the 'natural' rate of output is suboptimal because of various (unspecified) distortions. However, in any rational expectations equilibrium, inflation expectations are fulfilled and output equals its natural rate. If inflation is bad in itself (other things being equal), then the best rational expectations equilibrium features zero expected and actual inflation. This is the second best: the best equilibrium that can be achieved subject to the constraint of rational expectations. The first-best would be for output to be at some ideal level greater than its natural rate but with inflation maintained at zero. However, this ideal outcome is not consistent with rational expectations. Now suppose inflation policy can be revised after expectations are formed. Then the zero-inflation policy that gives rise to the best rational expectations equilibrium is not time consistent: when it is time to set inflation, the central bank would like to set an inflation rate above zero since a small rise in inflation has no first-order effect on the welfare cost of inflation but does have a first-order effect on the welfare cost of output being less than its ideal level. The result, under discretionary monetary policy, is a tendency for inflation to be above its desired level (inflation bias) with no (positive) effect on output or employment.

Overtaxation of Capital and Liquidity

Another early contribution is that of Fischer (1980), who discussed a situation where a fiscal authority decides on the levels of taxation on labour and capital and of public expenditure. There, the problem is that tomorrow's capital stock depends on today's investment. Meanwhile, today's investment depends on expectations about future capital income taxes. The result is that a government that sets taxes sequentially will typically overtax capital income. Similarly, Calvo (1978) described the time consistency problem in a monetary economy with money in the utility function. He found that, when lump-sum taxation is not available, the Friedman rule of optimal monetary policy is not time consistent; in general, the government wants to expand the money supply to relax the government budget constraint. This is because monetary expansion, like capital taxation, is distortionary only *ex ante;* it is the *expectation* of monetary expansion (and the consequent inflation) that leads people to economize on liquidity, a socially free resource.

Relation with Game Theoretic Concepts

The relationship between time consistency, time inconsistency and various concepts in game theory have been much discussed. A common view, but by no means a consensus, has emerged, asserting that the best way to think about the situation is that the (Ramsey 1927) optimal policy, 'second best' or 'commitment solution' (these phrases are used interchangeably) is an equilibrium of one game, the time-consistent policy the equilibrium of another. The first game lets the government move before time starts, choosing a time- and state-contingent policy for the indefinite future. Thereafter it does not move again. The second game has the government moving sequentially, setting policy in each period as it arrives. When the policies implied by these equilibria differ, then we say that the Ramsey policy is time inconsistent. On the other hand, any equilibrium of the second game is time-consistent by construction. This view of course leaves open what the correct solution concept is for these various games. Chari and Kehoe (1990) and several successors discuss the appropriate solution concept for the second type of game. From this literature has emerged the concept of 'sustainable equilibrium' which roughly corresponds to the sequential equilibrium concept of Kreps and Wilson (1982) but modified to apply to economies with one large agent and many 'small' agents. A recent formulation can be found in Phelan and Stacchetti (2001).

Solving the Time Inconsistency Problem

The literature on time consistency may usefully be divided into two parts: one attempts to characterize the equilibrium of the sequential-move game, the other tries to solve the time-inconsistency problem by somehow erasing the difference between the Ramsey policy and the time-consistent policy. A celebrated paper in the second category is Lucas and Stokey (1983), which discusses a dynamic monetary economy without capital. A price-taking representative agent chooses labour supply and the government sets labour taxes so as to minimize distortions subject to a government budget constraint. Government expenditure is exogenous. The government can issue state-contingent debt which it is committed to honouring. The main finding of the paper is that, if public debt has a sufficiently rich maturity structure, then the Ramsey optimal policy is time consistent. However, if only one-period state-contingent bonds are available, then the optimal policy is typically not time consistent. Persson et al. (1987) extend this result to monetary economies, showing how the government can render the optimal policy time consistent by accumulating nominal assets equal to the stock of money so that the outstanding stock of net nominal claims is zero. A minor mistake in that paper was pointed out by Calvo and Obstfeld (1990), but the basic result stands and applies quite generally, as explained in Alvarez et al. (2001). The latter paper establishes a link between the optimality of the Friedman rule (zero nominal interest rates)

and the possibility of rendering the Ramsey optimal policy time-consistent: for a wide class of economies they show that the Ramsey optimal policy can be made time-consistent if and only if the Friedman rule is optimal. Domínguez (2007) extends this result further for an economy with capital, showing that, if capital taxes are set one period in advance, then a sufficiently rich maturity structure of public debt is sufficient to render the optimal policy time consistent.

Another sub-literature in this category looks at reputational mechanisms that might render the optimal policy time consistent, or at least bring the time-consistent solution closer to the second-best optimum. In monetary policy, a key early contribution is Barro and Gordon (1983). In an environment with an expectations-augmented Phillips curve, it shows, using the well-known folk theorem from game theory, that the optimal monetary policy in the environment described by Kydland and Prescott (1977) can be sustained as a time-consistent equilibrium provided the policymaker is patient enough. A paper that analyses fiscal policy with a reputational-style approach is Kotlikoff et al. (1988). In this paper it is shown how the optimal tax scheme can be sustained in a two-period overlapping generations environment by threats of moving to the third-best equilibrium if any generation deviates.

Yet another set of solutions to the time-consistency problem of monetary policy is found in Rogoff (1985) and Persson and Tabellini (1993). The first, using an idea from industrial organization first published by Vickers (1985), shows that a monetary policymaker will typically be better off delegating monetary policy to a central banker that cares more about low inflation and less about output or employment than he or she does. Rogoff's result is aptly described as delegation to a 'conservative central banker'. This delegation improves welfare but does not achieve the second-best optimum. By contrast, the key result in Persson and Tabellini (1993) is that the second best can be achieved by signing a performance contract with the central banker.

Analysing What Happens When the Time Inconsistency Problem Cannot be Solved

The literature on characterizing time-consistent policy also divides into two parts: one focusing on a solution concept ('sustainable equilibrium') that is nearly always set-valued and the other on a refinement ('Markov perfect equilibrium') that often (but certainly not always) yields a unique equilibrium. The concept of Markov perfect equilibrium, whose purpose is essentially to rule out any reputational mechanisms, is defined in a game-theoretic setting in Maskin and Tirole (2001) and in a macroeconomic setting by, among several others, Klein et al. (2006). In the latter paper, the authors find that in the Markov perfect equilibrium labour tends to be under-taxed even in an environment where no other taxes are available and the only other endogenous variable is government spending. That is, a government acting sequentially tends to exaggerate the distortionary effects of taxation. This is in marked contrast to the case of capital and inflation taxes. The reason is that the current labour income tax *encourages* labour supply in the previous period, by intertemporal substitution. This effect is ignored by a sequentially moving government that thus neglects a beneficial effect of raising the current labour income tax rate.

Other papers studying Markov perfect equilibrium in a fiscal policy setting include Klein and Ríos-Rull (2003), who look at time-consistent capital and labour income taxation where capital taxes are set one period in advance and the budget has to balance in each period. The main finding is that a calibrated model can replicate the capital and labour taxes that we observe in, say, the United States, reasonably well. Krusell et al. (2004) consider public debt policy in an economy without capital and find that for positive initial debt there is a unique equilibrium but infinitely many steady states. For negative initial debt there are infinitely many equilibria, each associated with infinitely many steady states.

On the other hand, Phelan and Stacchetti (2001) study the set of sustainable equilibria in

an economy with capital but without public debt. The methods used are similar to those in Abreu et al. (1990). Fernandez-Villaverde and Tsyvinski (2002) consider all the sustainable equilibria in a stochastic environment with capital. They compare the best (from a welfare point of view) in that class with the Markov perfect equilibrium and the Ramsey equilibrium. Also, a literature is emerging on time-consistent policy under asymmetric information. Recent contributions include Sleet (2003) and Sleet and Yeltekin (2004).

A new departure in the study of time consistency of monetary policy is the consideration of the role of sticky prices. This introduces a new channel through which time inconsistency may arise: when prices are sticky and firms are bound to produce whatever is demanded at the given price, a surprise monetary expansion raises output. This is typically *ex post* welfare-improving in an economy suffering from some distortion, typically monopolistic rather than perfect competition. Important contributions include Albanesi et al. (2003a), who show that without commitment the economy can get stuck in an 'expectation trap' in the following sense. There are multiple, Pareto-ranked, equilibria. In the lower-ranked equilibria, the private sector expects high inflation. The measures the private sector takes to protect itself from high inflation create incentives for the policymaker to accommodate these expectations. On the other hand, in Albanesi et al. (2003b) the same authors show that optimal monetary policy is time consistent – and the Markov perfect equilibrium unique – in a wide class of models.

See Also

- ▶ Central Bank Independence
- ▶ Phillips Curve

Bibliography

Abreu, D., D. Pearce, and E. Stacchetti. 1990. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica* 58: 1041–1063.

Albanesi, S., V.V. Chari, and L. Christiano. 2003a. Expectation traps and monetary policy. *Review of Economic Studies* 70: 715–741.

Albanesi, S., V.V. Chari, and L. Christiano. 2003b. How severe is the time inconsistency problem in monetary policy? *Federal Reserve Bank of Minneapolis Quarterly Review* 27(3): 17–33.

Alvarez, F., P.J. Kehoe, and P.A. Neumayer. 2001. The time consistency of fiscal and monetary policies. Staff report no. 305, Federal Reserve Bank of Minneapolis.

Barro, R.J., and D.B. Gordon. 1983. Rules, discretion and reputation in a model of monetary policy. *Journal of Monetary Economics* 12: 101–121.

Calvo, G. 1978. On the time consistency of optimal policy in a monetary economy. *Econometrica* 46: 1411–1428.

Calvo, G.A., and M. Obstfeld. 1990. Time consistency of fiscal and monetary policy: A comment. *Econometrica* 58: 1245–1247.

Chari, V.V., and P.J. Kehoe. 1990. Sustainable plans. *Journal of Political Economy* 98: 784–802.

Domínguez, B. 2007. On the time-consistency of optimal capital taxes. *Journal of Monetary Economics* 54: 686–705.

Fernandez-Villaverde, J., and A. Tsyvinski. 2002. *Optimal fiscal policy in a business cycle model without commitment*. Mimeo: University of Pennsylvania.

Fischer, S. 1980. Dynamic inconsistency, cooperation and the benevolent dissembling government. *Journal of Economic Dynamics and Control* 2: 93–107.

Klein, P., P. Krusell, and J.-V. Ríos-Rull. 2006. *Time-consistent public expenditure*. Mimeo: Princeton University.

Klein, P., and J.-V. Ríos-Rull. 2003. Time-consistent optimal fiscal policy. *International Economic Review* 44: 1217–1245.

Kotlikoff, L.J., T. Persson, and L.E. Svensson. 1988. Social contracts as assets: A possible solution to the time-consistency problem. *American Economic Review* 78: 662–677.

Kreps, D.M., and R.B. Wilson. 1982. Sequential equilibria. *Econometrica* 50: 863–894.

Krusell, P., F. Martin, and J.V. Ríos-Rull. 2004. *On the determination of government debt*. Mimeo: Simon Fraser University.

Kydland, F.E., and E.C. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–491.

Lucas, R.E., and N.L. Stokey. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12: 55–93.

Maskin, E., and J. Tirole. 2001. Markov perfect equilibrium. *Journal of Economic Theory* 100: 191–219.

Persson, M., T. Persson, and L. Svensson. 1987. Time consistency of fiscal and monetary policy. *Econometrica* 55: 1419–1431.

Persson, T., and L.E. Svensson. 1989. Why a stubborn conservative would run a deficit: Policy with time-

inconsistent preferences. *Quarterly Journal of Economics* 104: 325–345.
Persson, T., and G. Tabellini. 1993. Designing institutions for monetary stability. *Carnegie-Rochester Conference Series on Public Policy* 39: 53–84.
Phelan, C., and E. Stacchetti. 2001. Sequential equilibria in a Ramsey tax model. *Econometrica* 69: 1491–1518.
Ramsey, F.P. 1927. A contribution to the theory of taxation. *Economic Journal* 37(145): 47–61.
Rogoff, K. 1985. The optimal degree of commitment to an intermediate monetary target. *Quarterly Journal of Economics* 100: 1169–1190.
Sleet, C. 2003. Optimal taxation with private government information. *Review of Economic Studies* 71: 1217–1239.
Sleet, C., and S. Yeltekin. 2004. *Credible social insurance*. Mimeo: University of Iowa.
Vickers, J. 1985. Delegation and the theory of the firm. *Economic Journal Supplement* 95: 138–147.

Time Preference

Murray N. Rothbard

Keywords

Austrian economics; Azpilcueta Navarrus, M. de; Bailey, S.; Böhm-Bawerk, E. von; Capital theory; Capitalization; Fetter, F. A.; Fisher, I.; Galiani, F.; Interest rates; Interest theory; Longfield, M.; Lottini da Volterra, G. F.; Marginal productivity; Menger, C.; Mises, L. E. von; Rae, J.; Summenhart, C.; Time preference; Turgot, A. R. J.; Usury; Value and distribution theory; Value theory

JEL Classifications

D9

Time preference is the insight that people prefer 'present goods' (goods available for use at present) to 'future goods' (present expectations of goods becoming available at some date in the future), and that the social rate of time preference, the result of the interactions of individual time preference schedules, will determine and be equal to the pure rate of interest in a society. The economy is pervaded by a time market for present as against future goods, not only in the market for loans (in which creditors trade present money for the right to receive money in the future), but also as a 'natural rate' in all processes of production. For capitalists pay out present money to buy or rent land, capital goods, and raw materials, and to hire labour (as well as buying labour outright in a system of slavery), thereby purchasing expectations of future revenue from the eventual sales of product. Long-run profit rates and rates of return on capital are therefore forms of interest rate. As businessmen seek to gain profits and avoid losses, the economy will tend toward a general equilibrium, in which all interest rates and rates of return will be equal, and hence there will be no pure entrepreneurial profits or losses.

In centuries of wrestling with the vexed question of the justification of interest, the Catholic scholastic philosophers arrived at highly sophisticated explanations and justifications of return on capital, including risk and the opportunity cost of profit forgone. But they had extreme difficulty with the interest on a riskless loan, and hence denounced all such interest as sinful and usurious.

Some of the later scholastics, however, in their more favourable view of usury, began to approach a time preference explanation of interest. During a comprehensive demolition of the standard arguments for the prohibition of usury in his *Treatise on Contracts* (1499), Conrad Summenhart (1465–1511), theologian at the University of Tübingen, used time preference to justify the purchase of a discounted debt, even if the debt be newly created. When someone pays $100 for the right to obtain $110 at a future date, the buyer (lender) doesn't profit usuriously from the loan because both he and the seller (borrower) value the future $110 as being worth $100 at the present time (Noonan 1957).

A half-century later, the distinguished Dominican canon lawyer and monetary theorist at the University of Salamanca, Martin de Azpilcueta Navarrus (1493–1586) clearly set forth the concept of time preference, but failed to apply it to a defence of usury. In his *Commentary on Usury*

(1556), Azpilcueta pointed out that a present good, such as money, will naturally be worth more on the market than future goods, that is, claims to money in the future. As Azpilcueta put it:

> a claim on something is worth less than the thing itself, and . . . it is plain that that which is not usable for a year is less valuable than something of the same quality which is usable at once. (Gordon 1975, p. 215)

At about the same time, the Italian humanist and politician Gian Francesco Lottini da Volterra, in his handbook of advice to princes, *Avvedimenti civili* (1574), discovered time preference. Unfortunately, Lottini also inaugurated the tradition of moralistically deploring time preference as an overestimation of a present that can be grasped immediately by the senses (Kauder 1965, pp. 19–22).

Two centuries later, the Neapolitan abbé, Ferdinando Galiani (1728–87), revived the rudiments of time-preference in his *Della Moneta* (1751) (Monroe 1924). Galiani pointed out that just as the exchange rate of two currencies equates the value of a present and a spatially distant money, so the rate of interest equates present with future, or temporally distant, money. What is being equated is not physical properties, but subjective values in the minds of individuals.

These scattered hints scarcely prepare one for the remarkable development of a full-scale time preference theory of interest by the French statesman, Anne Robert Jacques Turgot (1727–81), who, in a relatively few hastily written contributions, anticipated almost completely the later Austrian theory of capital and interest (Turgot 1977). In the course of a paper defending usury, Turgot asked: why are borrowers willing to pay an interest premium for the use of money? The focus should not be on the amount of metal repaid but on the usefulness of the money to the lender and borrower. In particular, Turgot compares the 'difference in usefulness which exists at the date of borrowing between a sum currently owned and an equal sum which is to be received at a distant date', and notes the well-known motto, 'a bird in the hand is better than two in the bush'. Since the sum of money owned now 'is preferable to the assurance of receiving a similar sum in one or several years' time', returning the same principal means that the lender 'gives the money and receives only an assurance'. Therefore, interest compensates for this difference in value by a sum proportionate to the length of the delay. Turgot added that what must be compared in a loan transaction is not the value of money lent with the value repaid, but rather the 'value of the *promise* of a sum of money compared to the value of money available now' (Turgot 1977, pp. 158–9).

In addition, Turgot was apparently the first to arrive at the concept of *capitalization*, a corollary to time preference, which holds that the present capital value of any durable good will tend to equal the sum of its expected annual rents, or returns, discounted by the market rate of time preference, or rate of interest.

Turgot also pioneered in analysing the relation between the quantity of money and interest rates. If an increased supply of money goes to low time-preference people, then the increased proportion of savings to consumption lowers time preference and hence interest rates fall while prices rise. But if an increased quantity goes into the hands of high time-preference people, the opposite would happen and interest rates would rise along with prices. Generally, over recent centuries, he noted, the spirit of thrift has been growing in Europe and hence time preference rates and interest rates have tended to fall.

One of the notable injustices in the historiography of economic thought was Böhm-Bawerk's brusque dismissal in 1884 of Turgot's anticipation of his own timepreference theory of interest as merely a 'land fructification theory' (Böhm-Bawerk, vol. 1, 1884–9). Partly this dismissal stemmed from Böhm's methodology of clearing the ground for his own positive theory of interest by demolishing, and hence sometimes doing injustice to, his own forerunners (Wicksell 1911, p. 177). The unfairness is particularly glaring in the case of Turgot, because we now know that in 1876, only eight years before the publication of his history of theories of interest, Böhm-Bawerk wrote a glowing tribute to Turgot's theory of

interest in an as yet unpublished paper in Karl Knies's seminar at the University of Heidelberg (Turgot 1977, pp. xxix–xxx).

In the course of his demolition of the Ricardo–James Mill labour theory of value on behalf of a subjective utility theory, Samuel Bailey (1825) clearly set forth the concept of time preference. Rebutting Mill's statement that time, as a 'mere abstract word', could not add to value, Bailey declared that 'we generally prefer a present pleasure or enjoyment to a distant one', and therefore prefer present goods to waiting for goods to arrive in the future. Bailey, however, did not go on to apply his insight to interest.

In the mid-1830s, the Irish economist Samuel Mountifort Longfield worked out the later Austrian theory of capital as performing the service for workers of supplying money at present instead of waiting for the future when the product will be sold. In turn the capitalist receives from the workers a time discount from their productivity. As Longfield put it, the capitalist

> pays the wages immediately, and in return receives the value of [the worker's] labour,... [which] is greater than the wages of that labour. The difference is the profit made by the capitalist for his advances . . . as it were, the discount which the labourer pays for prompt payment. (Longfield 1971)

The 'pre-Austrian' time analysis of capital and interest was most fully worked out, in the same year 1834, by the Scottish and Canadian eccentric John Rae (1786–1872). In the course of attempting an anti-Smithian defence of the protective tariff, Rae, in his *Some New Principles on the Subject of Political Economy* (1834), developed the Böhm-Bawerkian time analysis of capital, pointing out that investment lengthens the time involved in the processes of production. Rae noted that the capitalist must weigh the greater productivity of longer production processes against waiting for them to come to fruition. Capitalists will sacrifice present money for a greater return in the future, the difference – the interest return – reflecting the social rate of time preference. Rae saw that people's time preference rates reflect their cultural and psychological willingness to take a shorter or longer view of the future. His moral preferences were clearly with the low time-preference thrifty as against the high time-preference people who suffer from a 'defect of the imagination'. Rae's analysis had little impact on economics until resurrected at the turn of the 20th century, whereupon it was generously hailed in the later editions of Böhm-Bawerk's history of interest theories (Böhm-Bawerk, vol. 1, 1959).

Time preference, as a concept and as a foundation for the explanation of interest, has been an outstanding feature of the Austrian School of economics. Its founder, Carl Menger (1840–1921), enunciated the concept of time preference in 1871, pointing out that satisfying the immediate needs of life and health are necessarily prerequisites for satisfying more remote future needs. In addition, Menger declared, 'all experience teaches that we humans consider a present pleasure, or one expected in the near future, more important than one of the same intensity which is not expected to occur until some more distant time' (Wicksell 1924, p. 195; Menger 1871, pp. 153–4). But Menger never extended time preference from his value theory to a theory of interest; and when his follower Böhm-Bawerk did so, he peevishly deleted this discussion from the second edition of his *Principles of Economics* (Wicksell 1924, pp. 195–6).

Böhm-Bawerk's *Capital and Interest* (1884) is the *locus classicus* of the time preference theory of interest. In his first, historical volume, he demolished all other theories, in particular the productivity theory of interest; but five years later, in his *Positive Theory of Capital* (1889), Böhm brought back the productivity theory in an attempt to combine it with a time preference explanation of interest (Böhm-Bawerk, vols 1 and 2, 1959). In his 'three grounds' for the explanation of interest, time preference constituted two, and the greater productivity of longer processes of production the third, Böhm ironically placing greatest importance upon the third ground. Influenced strongly by Böhm-Bawerk, Irving Fisher increasingly took the same path of stressing the marginal productivity of capital as the main determinant of interest (Fisher 1907, 1930).

With the work of Böhm-Bawerk and Fisher, the modern theory of interest was set squarely on the path of placing time preference in a subordinate role in the explanation of interest, determining only the rate of consumer loans and the supply of consumer savings, while the alleged productivity of capital determines the more important demand for loans and for savings. Hence, modern interest theory fails to integrate interest on consumer loans and producers' returns into a coherent explanation.

In contrast, Frank A. Fetter, building on Böhm-Bawerk, completely discarded productivity as an explanation of interest and constructed an integrated theory of value and distribution in which interest is determined solely by time preference, while marginal productivity determines the 'rental prices' of the factors of production (Fetter 1915, 1977). In his outstanding critique of Böhm-Bawerk, Fetter pointed out a fundamental error of the third ground in trying to explain the return on capital as 'present goods' earning a return for their productivity in the future; instead, capital goods are *future* goods, since they are only valuable in the expectation of being used to produce goods that will be sold to the consumer at a future date (Fetter 1902). One way of seeing the fallacy of a productivity explanation of interest is to look at the typical practice of any current microeconomics text: after explaining marginal productivity as determining the demand curve for factors with wage rates on the *y*-axis, the textbook airily shifts to interest rates on the *y*-axis to illustrate the marginal productivity determination of interest. But the analog on the y-axis should not be interest, which is a ratio and not a price, but rather the *rental price* (price per unit time) of a capital good. Thus, interest remains totally unexplained. In short, as Fetter pointed out, marginal productivity determines rental prices, and time preference determines the rate of interest, while the capital value of a factor of production is the expected sum of future rents from a durable factor discounted by the rate of time preference or interest.

The leading economist adopting Fetter's pure time preference view of interest was Ludwig von Mises, in his *Human Action* (Mises 1949). Mises amended the theory in two important ways. First, he rid the concept of its moralistic tone, which had been continued by Böhm-Bawerk, implicitly criticizing people for 'under'-estimating the future. Mises made clear that a positive time preference rate is an essential attribute of human nature. Secondly, and as a corollary, whereas Fetter believed that people could have either positive or negative rates of time preference, Mises demonstrated that a positive rate is deducible from the fact of human action, since by the very nature of a goal or an end people wish to achieve that goal as soon as possible.

See Also

- ▶ Böhm-Bawerk, Eugen von (1851–1914)
- ▶ Fisher, Irving (1867–1947)

Bibliography

Bailey, S. 1825. *A critical dissertation on the nature, measure, and causes of value.* New York: Kelley, 1967.

Böhm-Bawerk, E. von. 1884–9. *Capital and interest,* vols. 1–2, 4th ed. South Holland: Libertarian Press, 1959.

Fetter, F.A. 1902. The 'roundabout process' in the interest theory. *Quarterly Journal of Economics* 17: 163–180. Repr. in Fetter (1977).

Fetter, F.A. 1915. *Economic principles,* vol. 1. New York: The Century.

Fetter, F.A. 1977. In *Capital, interest, and rent: Essays in the theory of distribution*, ed. M. Rothbard. Kansas City: Sheed Andrews and McMeel.

Fisher, I. 1907. *The rate of interest*. New York: Macmillan.

Fisher, I. 1930. *The theory of interest.* New York: Kelley & Millman, 1954.

Gordon, B. 1975. *Economic analysis before Adam Smith: Hesiod to Lessius*. New York: Barnes & Noble.

Kauder, E. 1965. *A history of Marginal utility theory.* Princeton: Princeton University Press.

Longfield, S.M. 1971. In *The economic writings of Mountifort Longfield*, ed. R.D.C. Black. Clifton: Kelley.

Menger, C. 1871. In *Principles of economics,* ed. J. Dingwall and B. Hoselitz. Glencoe: Free Press, 1950.

Mises, L. von. 1949. *Human action*: *A treatise on economics,* 3rd revised ed., Chicago: Regnery, 1966.

Monroe, A., ed. 1924. *Early economic thought*. Cambridge, MA: Harvard University Press.

Noonan, J.T. Jr. 1957. *The scholastic analysis of usury.* Cambridge, MA: Harvard University Press.

Rae, J. 1834. Some new principles on the subject of political economy. In *John Rae*: *Political economist,* ed. R.W. James. Toronto: University of Toronto Press, 1965.
Turgot, A.R.J. 1977. In *The economics of A.R.J. Turgot*, ed. P.D. Groenewegen. The Hague: Martinus Nijhoff.
Wicksell, K. 1911. Böhm-Bawerk's theory of interest. In *Selected papers on economic theory,* ed. E. Lindahl. Cambridge, MA: Harvard University Press, 1958.
Wicksell, K. 1924. The new edition of Menger's *Grundsatze.* In *Selected papers on economic theory*, ed. E. Lindahl. Cambridge, MA: Harvard University Press, 1958.

Time Series Analysis

Francis X. Diebold, Lutz Kilian and Marc Nerlove

Abstract

The analysis of economic time series is central to a wide range of applications, including business cycle measurement, financial risk management, policy analysis based on structural dynamic econometric models, and forecasting. This article provides an overview of the problems of specification, estimation and inference in linear stationary and ergodic time series models as well as non-stationary models, the prediction of future values of a time series and the extraction of its underlying components. Particular attention is devoted to recent advances in multiple time series modelling, the pitfalls and opportunities of working with highly persistent data, and models of nonlinear dependence.

Keywords

ARFIMA models; ARIMA models; ARMA models; ARMAX models; Autocovariance generating functions; Autoregressive conditional heteroskedasticity (ARCH); Band-Pass filter; Co-integration; Common factors; Cournot, A.; Ergodicity and non-ergodicity; Estimation; Factor model forecasts; Forecasting; Generalized autoregressive conditionally heteroskedastic (GARCH); Generalized method of moments; Granger, C.; Heteroskedasticity; Hodrick–Prescott (HP) filter; Inference; Jevons, W.; Kalman filter; Least squares; Linear processes; Long memory models; Markov chain methods; Maximum likelihood; Minimum mean-square error (MMSE) criterion; Multiple time series analysis; Noise; Nonlinear time series analysis; Optimal prediction and extraction theory; Prediction; Principal components analysis; Regime switching models; Seasonal adjustment; Simultaneous-equations model; Slutsky, E.; Smooth transition regression models; Specification; Spectral analysis; Spurious regressions; State-space methods; Stationarity; Stochastic volatility models; Structural vector autoregressions; Time domain analysis; Time series analysis; Unit root; Unobserved components models; Vector autoregressions; Volatility dynamics; Wiener–Kolmogorov theory; Wold decomposition theorem; Yule, G

JEL Classifications

C1

Any series of observations ordered along a single dimension, such as time, may be thought of as a time series. The emphasis in time series analysis is on studying the dependence among observations at different points in time. What distinguishes time series analysis from general multivariate analysis is precisely the temporal order imposed on the observations. Many economic variables, such as GNP and its components, price indices, sales, and stock returns are observed over time. In addition to being interested in the contemporaneous relationships among such variables, we are often concerned with relationships between their current and past values, that is, relationships over time.

The study of time series of, for example, astronomical observations predates recorded history. Early writers on economic subjects occasionally made explicit reference to astronomy as the source of their ideas. For example, Cournot (1838)

stressed that, as in astronomy, it is necessary to recognize the *secular* variations which are independent of the periodic variations. Similarly, Jevons (1884) remarked that his study of short-term fluctuations used the methods of astronomy and meteorology. During the 19th century interest in, and analysis of, social and economic time series evolved into a new field of study independent of developments in astronomy and meteorology (see Nerlove et al. 1979, pp. 1–21, for a historical survey).

Harmonic analysis is one of the earliest methods of analysing time series thought to exhibit some form of periodicity. In this type of analysis, the time series, or some simple transformation of it, is assumed to be the result of the superposition of sine and cosine waves of different frequencies. However, since summing a finite number of such strictly periodic functions always results in a perfectly periodic series, which is seldom observed in practice, one usually allows for an additive stochastic component, sometimes called 'noise'. Thus, an observer must confront the problem of searching for 'hidden periodicities' in the data, that is, the unknown frequencies and amplitudes of sinusoidal fluctuations hidden amidst noise. An early method for this purpose is *periodogram analysis*, suggested by Stokes (1879) and used by Schuster (1898) to analyse sunspot data and later by others, principally William Beveridge (1921, 1922), to analyse economic time series.

Spectral analysis is a modernized version of periodogram analysis modified to take account of the stochastic nature of the entire time series, not just the noise component. If it is assumed that economic time series are fully stochastic, it follows that the older periodogram technique is inappropriate and that considerable difficulties in the interpretation of the periodograms of economic series may be encountered.

At the time when harmonic analysis proved to be inadequate for the analysis of economic and social time series, another way of characterizing such series was suggested by the Russian statistician and economist, Eugen Slutsky (1927), and by the British statistician, G.U. Yule (1921, 1926, 1927). Slutsky and Yule showed that, if we begin with a series of purely random numbers and then take sums or differences, weighted or unweighted, of such numbers, the new series so produced has many of the apparent cyclic properties that were thought at the time to characterize economic and other time series. Such sums or differences of purely random numbers and sums or differences of the resulting series form the basis for the class of autoregressive moving-average (ARMA) processes which are used for modelling many kinds of time series. ARMA models are examples of time domain representations of time series. Although the latter may look very different from spectral representations of time series, there is a one-to-one mapping between time domain analysis and spectral analysis. Which approach is preferred in practice is a matter only of convenience. The choice is often determined by the transparency with which a given question can be answered. The remainder of this article explores these two complementary approaches to the analysis of economic time series.

Basic Theory

Stationarity and Ergodicity of Time Series Processes

Consider a random variable x_t where $t \in N$, the set of integers; the infinite vector $\{x_t, t \in N\}$ is called a discrete time series. Let M denote a subset of T consecutive elements of N. The distribution of the finite dimensional vector $\{x_t, t \in M\}$ is a well-defined multivariate distribution function, $F_M(\cdot)$. The time series $\{x_t, t \in N\}$ is said to be *strictly stationary* if, for any finite subset M of N and any integer τ, the distribution function of $\{x_t, t \in M+\tau\}$ is the same as the distribution function of $\{x_t, t \in M\}$. In other words, the joint distribution function of the finite vector of observations on x_t is invariant with respect to the origin from which time is measured. All the unconditional moments of the distribution function, if they exist, are independent of the index t; in particular,

$$E(x_t) = \mu$$

$$\gamma(\tau) = E[x_t - \mu][x_{t+\tau} - \mu], \tag{1}$$

where $\gamma(\tau)$ is the autocovariance function and depends only on the difference in indices, τ. Time-series processes for which (1) holds, but which are not necessarily strictly stationary according to the definition above, are said to be weakly *stationary*, *covariance stationary*, or *stationary to the second order*. Time-series processes for which $F_M(\cdot)$ is multivariate normal for any subset M of N are called *Gaussian processes*. For Gaussian processes covariance stationarity implies strict stationarity.

In practice, we usually observe only one realization of a finite subset of the time series of interest, corresponding to one of the many possible draws of length T from $F_M(\cdot)$. The question is whether the moments of x_t may be inferred from one such realization; for example, from the time averages of sums (or sums of products) of the observed values of a time series. If the process is what is known as *ergodic*, time averages of functions of the observations on the time series at T time points converge in mean square to the corresponding population expectations of x_t across alternative draws, as $T \to \infty$ (Priestley 1981, pp. 340–3; Doob 1953, p. 465). It is possible for a process to be stationary, yet not ergodic. Consider, for example, the process $x_t^{(i)} = \eta^{(i)} + \varepsilon_t$, where $x_t^{(i)}$ denotes the i*th* draw for observation x_t from the universe of all possible draws for x_t. Suppose that $\eta^{(i)} \sim N(0, \lambda^2)$ is the mean of the i*th* draw and that $\varepsilon_t \sim N(0, \sigma^2)$ is independent of $\eta^{(i)}$. This process is clearly stationary in that the probability limit of the ensemble average is zero, yet the time average $\sum_{t=1}^{T} x_t^{(i)}/T = \eta^{(i)} + \sum_{t=1}^{T} \varepsilon_t/T$ converges to $\eta^{(i)}$ rather than zero, thus violating ergodicity.

The Wold Decomposition and General Linear Processes

Let $\{\varepsilon_t\}$ be one element of a time series of serially uncorrelated, identically distributed random variables with zero mean and variance σ^2. Then the infinite, one-sided moving average (MA) process

$$x_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}, \tag{2}$$

where $b_0 = 1$ and $\sum_{j=0}^{\infty} b_j^2 < \infty$, is also a well-defined stationary process with mean 0 and variance $\sigma^2 \sum_0^{\infty} b_j^2$. Processes of this form and, more generally, processes based on an infinite two-sided MA of the same form are called *linear processes*, are always ergodic, and play a key role in time series analysis (Hannan 1970).

The importance of the process (2) is underscored by the Wold decomposition theorem (Wold 1938), which states that any weakly stationary process may be decomposed into two mutually uncorrelated component processes, one an infinite one-sided MA of the form (2) and the other a so-called linearly deterministic process, future values of which can be predicted exactly by some linear function of past observations. The linearly deterministic component is non-ergodic.

Linear Processes in Time and Frequency Domains

Autocovariance and Autocovariance Generating Functions

The autocovariance function of a stationary process, defined in (1) above, or its matrix generalization for vector processes, provides the basic representation of time dependence for weakly stationary processes. For the stationary process defined in (2), it is

$$\gamma(\tau) = \sigma^2 \sum_{j=0}^{\infty} b_j b_{j+\tau}. \tag{3}$$

Let z denote a complex scalar. Then the autocovariance generating transform is defined as

$$g(z) = \sum_{-\infty}^{\infty} \gamma(\tau) z^{\tau} \tag{4}$$

in whatever region of the complex plane the series on the right-hand side converges. If the series $\{x_t\}$ is covariance stationary, convergence will occur in an annulus about the unit circle. The autocovariance generating transform for the one-sided MA process defined in (2) is

$$g(z) = \sigma^2 B(z) B\left(z^{-1}\right) \tag{5}$$

where

$$B(z) = \sum_{k=0}^{\infty} b_k z^k.$$

If $B(z)$ has no zeros on the unit circle, the process defined in (2) is invertible and also has an infinite-order autoregressive (AR) representation as

$$A(L)x_t = \varepsilon_t, \tag{6}$$

where L is the lag operator such that $L^j x_t = x_{t-j}$ and $A(L) = a_0 + a_1 L + a_2 L^2 + \ldots$.

So-called ARMA processes have an autocovariance generating transform which is a rational function of z. If the ARMA process is both stationary and invertible, $g(z)$ may be written as

$$G(z) = \frac{P(z)P(z^{-1})}{\mathrm{Q}(z)\mathrm{Q}(z^{-1})} = \sigma^2 \frac{\prod_{k=1}^{m}(1-\beta_k z)(1-\beta_k z^{-1})}{\prod_{j=1}^{n}\left(1-\alpha_j z\right)\left(1-\alpha_j z^{-1}\right)} \tag{7}$$

where $|\beta_k|$, $|\alpha_j| < 1 \ \forall \ j$, k. Then the corresponding ARMA model is

$$Q(\mathrm{L})x_t = P(L)\varepsilon_t, \tag{8}$$

where

$$Q(\mathrm{L}) = \prod_{j=1}^{n}\left(1-\alpha_j L\right) \text{ and}$$

$$P(L) = \prod_{k=1}^{m}(1-\beta_k L).$$

Spectral Density Functions

If the value of z lies on the complex unit circle, it follows that $z = e^{-i\lambda}$, where $i = \sqrt{-1}$ and $-\pi \leq \lambda \leq \pi$. Substituting for z in the autocovariance generating transform (5) and dividing by 2π, we obtain the *spectral density function* of a linearly non-deterministic stationary process $\{x_t\}$ in terms of the frequency λ:

$$\begin{aligned} f(\lambda) &= (1/2\pi) g\left(e^{i\lambda}\right) \\ &= \left(\sigma^2/2\pi\right) B\left(\mathrm{e}^{i\lambda}\right) B\left(\mathrm{e}^{-i\lambda}\right) \\ &= (1/2\pi) \sum_{-\infty}^{\infty} \gamma(\tau) \mathrm{e}^{-i\lambda\tau}, \ -\pi \leq \lambda < \pi. \end{aligned} \tag{9}$$

Thus, the spectral density function is the Fourier transform of the autocovariance function. It can be shown that for a process with absolutely summable autocovariances the spectral density function exists and can be used to compute all of the autocovariances, so the same time series can be characterized equivalently in terms of the autocovariance function in the time-domain or in terms of the spectral density function in the frequency domain.

The spectral density function for a linearly non-deterministic, stationary, real-valued time series is a real-valued, non-negative function, symmetric about the origin, defined in the interval $[-\pi, \pi]$:

$$f(\lambda) = (1/2\pi)\left[\gamma(0) + 2\sum_{\tau=1}^{\infty}\gamma(\tau)\cos\lambda\tau\right]. \tag{10}$$

Moreover,

$$E(x_t - \mu)^2 = \int_{-\pi}^{\pi} f(\lambda)\mathrm{d}\lambda, \tag{11}$$

so that the spectral density function is a frequency-band decomposition of the variance of $\{x_t\}$.

When the process generating $\{x_t\}$ is merely stationary, that is, when $\{x_t\}$ may have a linearly deterministic component, the spectral density function is

$$f(\lambda) = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda\tau} dF(\lambda), \tag{12}$$

where $\mathrm{F}(\lambda)$ is a distribution function (Doob 1953, p. 488). Note that deterministic seasonal effects, for example, may cause a jump in the spectral distribution function.

The autocovariance function, its generating transform and the spectral distribution function

all have natural generalizations to the multivariate case, in which $\{x_t\}$ can be thought of as a vector of time-series processes.

The estimation and analysis of spectral density and distribution functions play an important role in all forms of time-series analysis. More detailed treatments are Doob (1953), Fishman (1969), Koopmans (1974), Fuller (1976), Nerlove et al. (1979, Ch. 3) and Priestley (1981).

Unobserved Components (UC) Models

In the statistical literature dealing with the analysis of economic time series it is common practice to classify the types of movements that characterize a time series as trend, cyclical, seasonal, and irregular components. The idea that a time series may best be viewed as being composed of several unobserved components is by no means universal, but it plays a fundamental role in many applications, for example, the choice of methods for seasonal adjustment. Nerlove et al. (1979, Ch. 1) review the history of the idea of unobserved components in economics from its origin early in the 19th century.

In the 1960s, Nerlove (1964, 1965, 1967) and Granger (1966) suggested that the typical spectral shape of many economic time series could be accounted for by the superposition of two or more independent components with specified properties. There are basically two approaches to the formulation of UC models: Theil and Wage (1964) and Nerlove and Wage (1964), Nerlove (1967) and Grether and Nerlove (1970) choose the form of components in such a way as to replicate the typical spectral shape of the series which represents their superposition. For example, let T_t represent the trend component, C_t the cyclical, S_t the seasonal, and I_t the irregular of a monthly time series; then the observed series can be represented as

$$y_t = T_t + C_t + S_t + I_t, \tag{13}$$

where

$$T_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p,$$

$$C_t = \frac{1 + \beta_1 L + \beta_2 L^2}{(1 - \alpha_1 L)(1 - \alpha_2 L)} \varepsilon_{1t},$$

$$S_t = \frac{1 + \beta_3 L + \beta_4 L^2}{1 - \gamma L^{12}} \varepsilon_{2t},$$

$$I_t = \varepsilon_{3t},$$

and ε_{1t}, ε_{2t}, and ε_{3t}, are i.i.d. normal variables with variances σ_{11}, σ_{22}, and σ_{33}, respectively. This approach has been carried forward by Harvey (1984), Harvey and Peters (1990) and Harvey and Todd (1984).

An alternative approach is to derive the components of the UC model from a well-fitting ARMA model (obtained after suitably transforming the data), given sufficient a priori identifying restrictions on the spectral properties of the components. See Box et al. (1978), Pierce (1978, 1979), Burman (1980), Hillmer and Tiao (1982), Hillmer et al. (1983), Bell and Hillmer (1984), Burridge and Wallis (1985), and Maravall (1981, 1984). The basis of this procedure is the fact that every stationary UC model, or the stationary part of every UC model, has an equivalent ARMA form, the so-called canonical form of the UC model (Nerlove and Wage 1964; Nerlove et al. 1979, Ch. 4).

Specification, Estimation, Inference and Prediction

Autocovariance and Spectral Density Functions

Suppose we have a finite number of observations of a realization of the process generating the time series, say $x_1, \ldots, x_T$. For expository purposes it is assumed that all deterministic components of x_t have been removed. If μ is unknown, this may be accomplished by subtracting the sample mean of the time series observations from the data prior to the analysis. For a zero mean series xt there are basically two ways of estimating $\gamma(\tau)$ defined in (1): the first is the biased estimator

$$c(\tau) = (1/T)\sum_{t=1}^{T-|\tau|} x_t x_{t+|\tau|}, \quad \tau = 0, \quad \pm 1, \ldots, \pm M, M \leq (T-1) \tag{14}$$

The second is the unbiased estimator

$$\tilde{c}(\tau) = [1/(T-|\tau|)]\sum_{1}^{T-|\tau|} x_t x_{t+|\tau|}, \quad \tau = 0, \pm 1, \ldots, \pm M, M \leq (T-1). \tag{15}$$

Although c(τ) is biased in finite samples, it is asymptotically unbiased. The key difference between c(τ) and $\tilde{c}(\tau)$ is that c(τ) is a positive definite function of τ whereas $\tilde{c}(\tau)$ is not (Parzen 1961, p. 981). The variance and covariances of the estimated autocovariances are derived, *inter alia*, by Hannan (1960), and Anderson (1971). As T→ ∞, both tend to zero, as the estimates are asymptotically uncorrelated and consistent. However,

$$E[c(\tau) - Ec(\tau)]^2 / E[c(\tau)] \to \infty \quad \text{as} \quad \tau/T \to 1. \tag{16}$$

This property accounts for the failure of the estimated autocorrelation function

$$r(\tau) = c(\tau)/c(0) \tag{17}$$

to damp down as $\tau \to \infty$, as it should for a stationary, linearly non-deterministic process (Hannan 1960, p. 43).

A 'natural' estimator of the spectral density function is obtained by replacing $\gamma(\tau)$ in (10) by c(τ) or $\tilde{c}(\tau)$. The resulting estimator is proportional, at each frequency, to a sample quantity called the periodogram:

$$I_T(\lambda) = (2/T)\left|\sum_{1}^{T} e^{i\lambda t} x_t\right|^2 \tag{18}$$

usually evaluated at the equi-spaced frequencies

$$\lambda = 2k\pi/T, \quad k = 1, 2, \ldots, [T/2] \tag{19}$$

in the interval [0, π]. Although, for a stationary, nonlinearly deterministic process, the periodogram ordinates are asymptotically unbiased estimates of the spectral densities at the corresponding frequencies, they are not consistent estimates; moreover, the correlation between adjacent periodogram ordinates tends to zero with increasing sample size. The result is that the periodogram presents a jagged appearance which is increasingly difficult to interpret as more data become available.

In order to obtain consistent estimates of the spectral density function at specific frequencies, it is common practice to weight the periodogram ordinates over the frequency range or to form weighted averages of the autocovariances at different lags. There is a substantial literature on the subject. The weights are called a 'spectral window'. Essentially the idea is to reduce the variance of the estimate of an average spectral density around a particular frequency by averaging periodogram ordinates which are asymptotically unbiased and independently distributed estimates of the corresponding ordinates of the spectral density function. Related weights can also be applied to the estimated autocovariances which are substituted in (10); this weighting system is called a 'lag window'.

Naturally the sampling properties of the spectral estimates depend on the nature of the 'window' used to obtain consistency (see Priestley 1981, pp. 432–94 for further discussion). Regardless of the choice of window, the 'bandwidth' used in constructing the window must decrease at a suitable rate as the sample size grows. In the spectral window approach, this means that the window width must decrease at a slower rate than the sample size. In the lag window approach, this means that the number of included autocovariances must increase at a slower rate than the sample size.

ARMA Models

The autocovariance function and the spectral density function for a time series represent nonparametric approaches to describing the data. An alternative approach is to specify and estimate a parametric ARMA model for x_t This approach

involves choosing the orders of the polynomials P and Q in (7) and (8) and perhaps also specifying that one or more coefficients are zero or placing other restrictions on P and Q. The problem then becomes one of estimating the parameters of the model.

Despite the poor statistical properties of the estimated autocovariance function and a related function called the partial autocorrelation function, these are sometimes used to specify the orders of the polynomials P and Q. An alternative approach is to select the model that minimizes the value of information-theoretic criteria of the form

$$\mathrm{IC}(i) = \log\left(\widehat{\sigma}_i^2\right) + k_i c_T, \tag{20}$$

where k_i refers to the number of estimated parameters in the candidate models $i = 1, \ldots, M$, and $\widehat{\sigma}_i^2$ to the corresponding maximum likelihood estimate of the residual variance. Such criteria incorporate a trade-off between the fit of a model and its degree of parsimony. That trade-off depends on the penalty term c_t (Akaike 1970, 1974; Schwarz 1978). There is no universally accepted choice for c_T. For $c_T = 2/T$ expression (20) reduces to the Akaike information criterion (AIC), for example, and for $c_T = ln(T)/T$ to the Schwarz information criterion (SIC). The asymptotic properties of alternative criteria will depend on the objective of the user and the class of models considered.

Given the orders of the AR and MA components, a variety of maximum likelihood or approximate maximum likelihood methods are available to estimate the model parameters. Newbold (1974) shows that, if x_t is characterized by (8) with $\varepsilon_t \sim NID(0, \sigma^2)$, then the exact likelihood function for the parameters of $P(\cdot)$ and $Q(\cdot)$ is such that the maximum likelihood estimates of the parameters and the least-squares (LS) estimates (in general highly nonlinear) are asymptotically identical. Only in the case of a pure AR model are the estimates linear conditional on the initial observations. Several approximations have been discussed (Box and Jenkins 1970; Granger and Newbold 1977; Nerlove et al. 1979, pp. 121–5).

Exact maximum likelihood estimation of ARMA models has been discussed by, inter alia, Newbold (1974), Anderson (1977), Ansley (1979), and Harvey (1981). Following Schweppe (1965), Harvey suggests the use of the Kalman filter to obtain the value of the exact-likelihood function, which may be maximized by numerical methods. The Kalman filter approach is easily adapted to the estimation of UC models in the time domain.

An alternative to exact or approximate maximum-likelihood estimation in the time domain was suggested by Hannan (1969). Estimates may be obtained by maximizing an approximate likelihood function based on the asymptotic distribution of the periodogram ordinates defined in (18). These are asymptotically independently distributed (Brillinger 1975, p. 95), and the random variables $2I_t(\lambda)/f(\lambda)$ have an asymptotic χ^2 distribution with two degrees of freedom (Koopmans 1974, pp. 260–5). This means that the asymptotic distribution of the observations, $\{x_1, \ldots, x_T\}$ is proportional to

$$\prod_{j=0}^{[T/2]} \left[1/f\left(\lambda_j\right)\right] \exp\left[-I\left(\lambda_j\right)/f\left(\lambda_j\right)\right] \tag{21}$$

where $\lambda_j = 2j\pi/T$, $j = 0, \ldots, [T/2]$, are the equi-spaced frequencies in the interval $[0, \pi]$ at which the periodogram is evaluated (Nerlove et al. 1979, pp. 132–6). Since the true spectral density $f(\lambda)$ depends on the parameters characterizing the process, this asymptotic distribution may be interpreted as a likelihood function. Frequency domain methods, as these are called, may easily be applied in the case of UC models.

Whether approximate or exact maximum-likelihood estimation methods are employed, inference may be based on the usual criteria related to the likelihood function. Unfortunately, serious difficulties may be encountered in applying the asymptotic theory, since the small sample distribution of the maximum likelihood estimator may differ greatly from the limiting distribution in important cases (Sargan and Bhargava 1983; Anderson and Takemura 1986).

Prediction and Extraction

The problem of prediction is essentially the estimation of an unknown future value of the time series itself; the problem of extraction, best viewed in the context of UC models described in section "Unobserved Components (UC) Models", is to estimate the value of one of the unobserved components at a particular point in time, not necessarily in the future. Problems of trend extraction and seasonal adjustment may be viewed in this way (Grether and Nerlove 1970). How the prediction (or extraction) problem is approached depends on whether we are assumed to have an infinite past history and, if not, whether the parameters of the process generating the time series are assumed to be known. In practice, of course, an infinite past history is never available, but a very long history is nearly equivalent if the process is stationary or can be transformed to stationarity. It is common, as well, to restrict attention to linear predictors, which involves no loss of generality if the processes considered are Gaussian and little loss if merely linear. To devise a theory of optimal prediction or extraction requires some criterion by which to measure the accuracy of a particular candidate. The most common choice is the minimum mean-square error (MMSE) criterion, which is also the conditional expectation of the unknown quantity. For a discussion of alternative loss functions see Granger (1969) and Christoffersen and Diebold (1996, 1997).

The theory of optimal prediction and extraction due to Kolmogorov (1941) and Wiener (1949) and elaborated by Whittle (1963) for discrete processes assumes a possibly infinite past history and known parameters. As a special case of the Wiener–Kolmogorov theory for non-deterministic, stationary processes, consider the linear process defined by (2). Since the ε_t are i.i.d. with zero mean and variance σ^2, it is apparent that the conditional expectation of x_{t+v}, given all innovations from the infinite past to t, is

$$\widehat{x}_{t+v} = b_v\varepsilon_t + b_{v+1}\varepsilon_{t+1} + \ldots \qquad (22)$$

Of course, even if the parameters b_j, $j = 0, 1, \ldots,$ are assumed to be known, the series $\{\varepsilon_t\}$ is not directly observable. The ε_t 's are sometimes called the *innovations* of the process, since it is easy to show that $\varepsilon_{t+1} = x_{t+1} - \widehat{x}_{t+1}$ is the one-step ahead prediction error. If the process is invertible, it has the autoregressive representation (6) and so can be expressed solely in terms of the, generally infinite-order, autoregression

$$\widehat{x}_{t+v} = D(L)x_t, \qquad (23)$$

where the generating transform of the coefficients of D is

$$D(z) = \frac{1}{B(z)}\left[\frac{B(z)}{z^v}\right]_+.$$

The operator $[\cdot]_+$ eliminates terms involving negative powers of z.

The problem of extraction is best viewed in the context of multiple time series; in general we wish to 'predict' one time series $\{y_t\}$ from another related series $\{x_t\}$. It is not necessary that the series $\{y_t\}$ actually be observed as long as its relationship to an observed series $\{x_t\}$ can be described (Nerlove et al. 1979, Ch. 5).

The Kalman filter approach to prediction and extraction (Kalman 1960) is both more special and more general than the Wiener–Kolmogorov theory: attention is restricted to finite-dimensional parameter spaces and linear processes, but these processes need not be stationary. The parameters may vary with time, and we do not require an infinite past. This approach represents a powerful tool of practical time-series analysis and may be easily extended to multiple time series. A full discussion, however, requires a discussion of 'state-space representation' of time series processes and is beyond the scope of this entry (Harvey 1989).

Multiple Time Series Analysis

A general treatment of multiple time series analysis is contained in Hannan (1970). The two-variable case will serve to illustrate the matter. Two stationary time series $\{x_t\}$ and $\{y_t\}$ are said to be jointly stationary if their joint distribution function does not depend on the origin from

which time is measured. Joint stationarity implies, but is not in general implied by, weak or covariance joint stationarity; that is, $cov(x_t, y_s)$ is a function of $s - t$ only. In this case the cross-covariance function is

$$\gamma_{yx}(\tau) = \mathrm{E}\left[y_t - \mu_y\right]\left[x_{t-\tau} - \mu_x\right], \tag{24}$$

where $\mu_x = Ex_t$ and $\mu_y = Ey_t$. Note that $\gamma_{yx}(\tau)$ and $\gamma_{xy}(\tau)$ are, in general, different. The cross-covariance generating function is defined as

$$g_{yx}(z) = \sum_{-\infty}^{\infty} \gamma_{yx}(\tau) z^{\tau} \tag{25}$$

in that region of the complex plane in which the right-hand side of (25) converges. For two jointly stationary series this occurs in an annulus containing the unit circle. In this case, the cross-spectral density function is defined as

$$f_{yx}(\lambda) = (1/2\pi) g_{yx}\left(\mathrm{e}^{i\lambda}\right). \tag{26}$$

Since $\gamma_{yx}(\tau)$ and $\gamma_{xy}(\tau)$ are not equal, the cross-spectral density function is complex valued and can be decomposed into a real part (the co-spectral density) and a complex part (the quadrature spectral density):

$$f_{yx}(\lambda) = c_{yx}(\lambda) + i q_{yx}(\lambda). \tag{27}$$

In polar form, the cross-spectral density may be written as

$$f_{yx}(\lambda) = \alpha_{yx}(\lambda) \exp\left[i\varphi_{yx}(\lambda)\right], \tag{28}$$

where $\alpha_{yx}(\lambda) = \left[c_{yx}^2(\lambda) + q_{yx}^2(\lambda)\right]^{1/2}$ is called the amplitude or gain, and where $\varphi_{yx}(\lambda) = \arctan\{-q_{yx}(\lambda)/c_{yx}(\lambda)\}$ is called the phase. Another useful magnitude is the coherence between the two series, defined as

$$\rho_{yx}(\lambda) = \frac{\left|f_{yx}(\lambda)\right|^2}{f_{xx}(\lambda) f_{yy}(\lambda)}, \tag{29}$$

which measures the squared correlation between y and x at a frequency λ. Clearly, $\rho_{yx}(\lambda) = \rho_{xy}(\lambda)$. Estimation of cross-spectral density functions and related quantities is discussed in Priestley (1981, pp. 692–712).

Often it is convenient to impose additional parametric structure in modelling multiple time series. The workhorse multiple time series model in econometrics has been the covariance-stationary K-dimensional vector autoregressive model, which may be viewed as a natural generalization of the univariate AR model discussed earlier:

$$A(L)x_t = \varepsilon_t \tag{30}$$

where $A(L) = I_K - A_1 L - \ldots - A_p L^p$. Here each variable in x_t is regressed on its own lags as well as lags of all other variables in x_t up to some prespecified lag order p. This *vector autoregression* (VAR) can also be viewed as an approximation to a general linear process x_t, and may be estimated by LS.

Similarly, the formulation of ARMA and UC models discussed earlier may be extended to the multivariate case by interpreting the polynomials in the lag operator as matrix polynomials and by replacing the scalar random variables by vectors. Although these vector ARMA and UC models bear a superficial resemblance to the corresponding univariate ones, their structure is, in fact, much more complicated and gives rise to difficult identification problems. In the univariate case, we can formulate simple conditions under which a given covariance function identifies a unique ARMA or UC model, but in the multivariate case these conditions are no longer sufficient. Hannan (1970, 1971) gives a complete treatment. State-space methods have also been employed to study the structure of multivariate ARMA models (Hannan 1976; and, especially, 1979).

Unit Roots, Co-integration and Long Memory

Standard tools for time series analysis have been developed for processes that are covariance stationary or have been suitably transformed to achieve covariance stationarity by removing (or explicitly

modelling) deterministic trends, structural breaks, and seasonal effects. The presence of a *unit root* in the autoregressive lag order polynomial of an ARMA process also violates the assumption of stationarity. Processes with a unit root are also called integrated of order one (or $I(1)$ for short) because they become covariance-stationary only upon being differenced once. In general, $I(d)$ processes must be differenced d times to render the process covariance-stationary.

The presence of unit roots has important implications for estimation and inference. When the scalar process x_t is $I(1)$ the variance of x_t will be unbounded, model innovations will have permanent effects on the level of x_t, the autocorrelation function does not die out, and x_t will not revert to a long-run mean. Moreover, coefficients of $I(1)$ regressors will have nonstandard asymptotic distributions, invalidating standard tools of inference.

The simplest example of an autoregressive integrated moving-average (ARIMA) process is the random walk process: $x_t = x_{t-1} + \varepsilon_t$. The potential pitfalls of regression analysis with $I(1)$ data are best illustrated by the problem of regressing one independent random walk on another. In that case, it can be shown that R^2 and $\widehat{\beta}$ will be random and that the usual t-statistic will diverge, giving rise to seemingly significant correlations between variables that are unrelated by construction. This *spurious regression* problem was first discussed by Yule (1926), further illustrated by Granger and Newbold (1974), and formally analyzed by Phillips (1986) and Phillips and Durlauf (1986). Similar problems arise in deterministically detrending $I(1)$ series (Nelson and Kang 1981; Durlauf and Phillips 1988). Unbalanced regressions, that is, regressions in which the regressand is not of the same order of integration as the regressor, may also result in spurious inference. An exception to this rule is inference on coefficients of mean zero $I(0)$ variables in regressions that include a constant term (Sims et al. 1991).

The standard response to dealing with $I(1)$ data is to difference the data prior to the analysis. There is one important exception to this rule. There are situations in which several variables are individually $I(1)$, but share a common unit root component. In that case, a linear combination of these variables will be $I(0)$:

$$c' x_t = u_t \sim I(0), c \neq 0 \tag{31}$$

where x_t denotes a K-dimensional vector of $I(1)$ variables and c is a $(K \times 1)$ parameter vector. In other words, these variables share a common stochastic trend. This phenomenon is known as *co-integration* (Granger 1981; Engle and Granger 1987) and c is known as the co-integrating vector. Clearly, c is not unique. It is common to normalize one element of c to unity. The LS estimator of c in (31) is consistent, but corrections for omitted dynamics are recommended (Stock and Watson 1993; Phillips and Hansen 1990). Co-integrating relationships have been used extensively in modelling long-run equilibrium relationships in economic data (Engle and Granger 1991).

Variables that are co-integrated are linked by an *error correction* mechanism that prevents the integrated variables from drifting apart without bound. Specifically, by the Granger representation theorem of Engle and Granger (1987), under some regularity conditions, any K-dimensional vector of co-integrated variables x_t can be represented as a vector error correction (VEC) model of the form:

$$\Delta x_t = \sum_{i=1}^{p-1} \Gamma_i \Delta x_{t-i} - \prod x_{t-p} \tag{32}$$

where Γ_i, $i = 1, \ldots, p-1$, and $\Pi \equiv BC$ are conformable coefficient matrices and Δ denotes the first-difference operator. Model (32) allows for up to r co-integrating relationships where r is the rank of Π. For $r = 0$; the error correction term in model (32) drops out and the model reduces to a difference-stationary VAR. For $r = K$, all variables are $I(0)$ and model (32) is equivalent to a stationary VAR in levels. Otherwise, there are $0 < r < K$ common trends. If the $(r \times K)$ matrix of co-integrating vectors, C, is known, the model in (32) reduces to

$$\Delta x_t = \sum_{i=1}^{p-1} \Gamma_i \Delta x_{t-i} - B z_{t-p} \tag{32}$$

T

Where $z_{t-p} \equiv Cx_{t-p}$, and the model may be estimated by LS; if only the rank r is known, the VEC model in (32) is commonly estimated by full information maximum likelihood methods (Johansen 1995).

Starting with Nelson and Plosser (1982), a large literature has dealt with the problem of statistically discriminating between $I(1)$ and $I(0)$ models for economic data. Notwithstanding these efforts, it has remained difficult to detect reliably the existence of a unit root (or of co-integration). The problem is that in small samples highly persistent, yet stationary processes are observationally equivalent to exact unit root processes. It may seem that not much could hinge on this distinction then, but it can be shown that $I(1)$ and $I(0)$ specifications that fit the data about equally well may have very different statistical properties and economic implications (Rudebusch 1993).

For processes with roots near unity in many cases neither the traditional asymptotic theory for $I(0)$ processes nor the alternative asymptotic theory for exact $I(1)$ processes will provide a good small-sample approximation to the distribution of estimators and test statistics. An alternative approach is to model the dominant root, ρ, of the autoregressive lag order polynomial as *local-to-unity* in the sense that $\rho = 1 - c/T$, $c > 0$. This asymptotic thought experiment gives rise to an alternative asymptotic approximation that in many cases provides a better small-sample approximation than imposing the order of integration or relying on unit root pretests (Stock 1991; Elliott 1998).

Stationary ARMA processes are 'short memory' processes in that their autocorrelation function dies out quickly. For large τ, ARMA autocorrelations decay approximately geometrically, that is $\rho(\tau) \approx r^{\tau}$, where r is a constant such that $|r| < 1$: In many applied contexts including volatility dynamics in asset returns, there is evidence that the autocorrelation function dies out much more slowly. This observation has motivated the development of the class of *fractionally integrated* ARMA (ARFIMA) models:

$$Q(L)(1-L)^{d}x_t = P(L)\varepsilon_t \qquad (33)$$

where d is a real number, as opposed to an integer (Baillie 1996). Stationarity and invertibility require $|d| < 0{:}5$, which can always be achieved by taking a suitable number of differences. The autocorrelation function of an ARFIMA process decays at a hyperbolic rate. For large τ, we have $\rho(\tau) \approx \tau^{2d-1}$, where $d < 1/2$ and $d \neq 0$. Such '*long memory*' models may be estimated by the two-step procedure of Geweke and Porter-Hudak (1983) or by maximum likelihood (Sowell 1992; Baillie et al. 1996). A detailed discussion including extensions to the notion of fractional co-integration is provided by Baillie (1996). Long memory may arise, for example, from infrequent stochastic regime changes (Diebold and Inoue 2001) or from the aggregation of economic data (Granger 1980; Chambers 1998). Perhaps the most successful application of long-memory processes in economics has been work on modelling the volatility of asset prices and powers of asset returns, yielding new insights into the behaviour of markets and the pricing of financial risk.

Nonlinear Time Series Models

The behaviour of many economic time series appears to change distinctly at irregular intervals, consistent with economic models that suggest the existence of floors and ceilings, buffer stocks and regime switches in the data. This observation has given rise to a large literature dealing with nonlinear time series models. Nonlinear time series models still have a Wold representation with linearly unpredictable innovations, but these innovations are nevertheless dependent over time. This has important implications for forecasting and for the dynamic properties of the model. For example, the effects of innovations in nonlinear models will depend on the path of the time series and the size of the innovation, and may be asymmetric.

Nonlinear Dynamics in the Conditional Mean

The increasing importance of nonlinear time series models in econometrics is best illustrated by two examples: hidden Markov chain models and smooth transition regression models of the conditional mean.

The idea of hidden Markov chains first attracted attention in econometrics in the context of *regime switching models* (Hamilton 1989). The original motivation was that many economic time series appear to follow a different process during recession phases of the business cycle than during economic expansions. This type of regime-switching behaviour may be modelled in terms of an unobserved discrete-valued state variable (for example, 1 for a recession and 0 for an expansion) that is driven by a Markov chain. The transition from one state to another is governed by a matrix of transition probabilities that may be estimated from past data. The essence of this method thus is that the future will in some sense be like the past. A simple example of this idea is the regime-switching AR(1) model:

$$x_t = a_{1st}x_{t-1} + \varepsilon_t, \varepsilon_t \sim NID\left(0, \sigma^2\right) \tag{34}$$

where the regime s_t is the outcome of an unobserved two-state Markov chain with st independent of $\varepsilon\tau$ for all t and τ. In this model, the time-varying slope parameter will take on different values depending on the state s. Once the model has been estimated by maximum likelihood methods, it is possible to infer how likely a given regime is to have generated the observed data at date t. An excellent review of the literature on hidden Markov models is provided by Cappé et al. (2005); for a general treatment of state space representations of nonlinear models, also see Durbin and Koopman (2001).

The idea of *smooth transition regression models* is based on the observation that many economic variables are sluggish and will not move until some state variable exceeds a certain threshold. For example, price arbitrage in markets will only set in once the expected profit of a trade exceeds the transaction cost. This observation has led to the development of models with fixed thresholds that depend on some observable state variable. Smooth transition models allow for the possibility that this transition occurs not all of a sudden at a fixed threshold but gradually, as one would expect in time series data that have been aggregated across many market participants. A simple example is the smooth-transition AR (1) model:

$$\begin{aligned} x_t &= \Phi(z_{t-1}, \ldots, z_{t-d}, \Gamma)x_{t-1} + \varepsilon_t \\ \varepsilon_t &\sim NID\left(0, \sigma^2\right) \end{aligned} \tag{35}$$

where $\Phi(\cdot)$ denotes the transition function, z_t is a zero mean state variable denoting the current deviation of x_t from a (possibly time-varying) equilibrium level and Γ is the vector of transition parameters. Common choices for the transition function are the logistic or the exponential function. For example, we may specify $\Phi(\cdot) = (\exp\{\gamma(z_{t-1})^2\})$ with $\gamma < 0$. If $z_{t-1} = 0$, $\Phi(\cdot) = 1$ and the model in (35) reduces to a random walk model; otherwise, $\Phi(\cdot) < 1$ and the model in (35) reduces to a stationary AR(1). The degree of mean reversion is increasing in the deviation from equilibrium. For further discussion see Granger and Teräsvirta (1993).

Nonlinear Dynamics in the Conditional Variance

While the preceding examples focused on nonlinear dynamics in the conditional mean, nonlinearities may also arise in higher moments. The leading example is the conditional variance. Many economic and financial time series are characterized by volatility clustering. Often interest centres on predicting these *volatility dynamics* rather than the conditional mean. The basic idea of modelling and forecasting volatility was set out in Engle's (1982) path-breaking paper on autoregressive conditional heteroskedasticity (ARCH). Subsequently, Bollerslev (1986) introduced the class of generalized autoregressive conditionally heteroskedastic (GARCH). Consider a decomposition of xt into the one-step ahead conditional mean $u_{t|t-1} \equiv E(x_t|\ \Omega_{t-1})$, and conditional variance $\sigma^2_{t|t-1} \equiv \mathrm{var}(x_t|\ \Omega_{t-1})$, where Ω_{t-1} denotes the information set at $t-1$:

$$x_t = u_{t|t-1} + \sigma_{t|t-1}v_t v_t \sim NID(0,1) \tag{36}$$

T

The leading example of a GARCH model of the conditional variance is the GARCH(1,1) model, which is defined by the recursive relationship

$$\sigma^2_{t|t-1} = \omega + \alpha\varepsilon^2_{t|t-1} + \beta\sigma^2_{t-1|t-2} \qquad (37)$$

where $\varepsilon_t \equiv \sigma_{t|t-1}v_t$, and the parameter restrictions $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ ensure that the conditional variance remains positive for all realizations of v_t. The standard estimation method is maximum likelihood. The basic GARCH(1,1) model may be extended to include higher-order lags, to allow the distribution of v_t to have fat tails, to allow for asymmetries in the volatility dynamics, to permit the conditional variance to affect the conditional mean, and to allow volatility shocks to have permanent effects or volatility to have long memory. It may also be extended to the multivariate case.

It follows directly from the formulation of the GARCH(1,1) model that the optimal, in the MMSE sense, one-step-ahead forecast equals $\sigma^2_{t+1|t}$. Similar expressions for longer horizons may be obtained by recursive updating. There is a direct link from the arrival of news to volatility measures and from volatility forecasts to risk assessments. These and alternative volatility models and the uses of volatility forecasts are surveyed in Andersen et al. (2006b). For a comparison of GARCH models with the related and complementary class of stochastic volatility models, see Andersen et al. (2006a) and Shephard (2005).

Applications

Time series analytic methods have many applications in economics. Here we consider five: (1) analysis of the cyclic properties of economic time series, (2) description of seasonality and seasonal adjustment, (3) forecasting, (4) dynamic econometric modelling, and (5) structural vector autoregressions.

Analysis of the Cyclic Properties of Economic Time Series

Suppose that the time series $\{x_t\}$ is a linearly nondeterministic stationary series and that the series $\{y_t\}$ is formed from $\{x_t\}$ by the linear operator

$$y_t = \sum_{j=m}^{n} w_j x_{t-j}, \qquad \sum_{m}^{n} w_j^2 > \infty. \qquad (38)$$

Such an operator is called a time-invariant linear filter. Analysis of the properties of such filters plays an important role in time series analysis since many methods of trend estimation or removal and seasonal adjustment may be represented or approximated by such filters. An interesting example that illustrates the potential pitfalls of using such filters is provided by Adelman (1965), who showed that the 20-year long swings in various economic series found by Kuznets (1961) may well have been the result of the trend filtering operations used in preliminary processing of the data. For a fuller treatment see Nerlove et al. (1979, pp. 53–7).

Since the 1980s, there has been increased interest in the use of nonlinear filters for extracting the business cycle component of macroeconomic time series. Examples include the band-pass filter (Christiano and Fitzgerald 2003) and the Hodrick–Prescott (HP) filter (Hodrick and Prescott 1997; Ravn and Uhlig 2002). The latter approach postulates that $y_t = \tau_t + c_t$, where τ_t denotes the trend component and ct the deviation from trend or 'cyclical' component of the time series y_t. The trend component is chosen to minimize the loss function:

$$\sum_{t=1}^{T} c_t^2 + \lambda \sum_{t=1}^{T} [(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2 \qquad (39)$$

where $c_t = y_t - \tau_t$ and λ is a pre-specified parameter that depends on the frequency of the observations. The trade-off in this optimization problem is between the degree to which the trend component fits the data and the smoothness of the trend.

Description of Seasonality and Seasonal Adjustment

Many economic time series exhibit fluctuations which are periodic within a year or a fraction thereof. The proper treatment of such seasonality, whether stochastic or deterministic, is the subject of a large literature, summarized rather selectively

in Nerlove et al. (1979, Ch. 1). More recent treatments can be found in Hylleberg (1992), Franses (1996) and Ghysels and Osborn (2001).

Seasonality may be modelled and its presence detected using spectral analysis (Nerlove 1964) or using time domain methods. Deterministic seasonality, in the form of model parameters that vary deterministically with the season, offers no great conceptual problems but many practical ones. Stochastic seasonality is often modelled in the form of seasonal unit roots. In that case, seasonal differencing of the data removes the unit root component. Multiple time series may exhibit seasonal cointegration. Sometimes it is convenient to specify stochastic seasonality in the form of an UC model (Grether and Nerlove 1970). Appropriate UC models may be determined directly or by fitting an ARIMA model and deriving a related UC model by imposing sufficient a priori restrictions (Hillmer and Tiao 1982; Bell and Hillmer 1984).

Forecasting

One of the simplest forecasting procedures for time series is exponential smoothing based on the relationship

$$\widehat{x}_{t+1|t} = (1-\theta)x + \theta\widehat{x}_{t|t-1} \tag{40}$$

where x_t, is the observed series and $\widehat{x}_{j|k}$ is the forecast of the series at time j made on the basis of information available up to time k. Muth (1960) showed that (40) provides an MMSE forecast if the model generating the time series is $x_t - x_{t-1} = \varepsilon_t - \theta\varepsilon_{t-1}$. Holt (1957) and Winters (1960) generalized the exponential smoothing approach to models containing more complex trend and seasonal components. Further generalization and proofs of optimality are contained in Theil and Wage (1964) and Nerlove and Wage (1964).

Perhaps the most popular approach to forecasting time series is based on ARIMA models of time series processes (Box and Jenkins 1970). The developments discussed in the preceding paragraph led to the development of UC models, which give rise to restricted ARIMA model forms (Nerlove et al. 1979). State-space representations of these models permit the application of the Kalman filter to both estimation and forecasting. Harvey (1984) presents a unified synthesis of the various methods.

More recently, the focus has shifted from traditional forecasting methods towards methods that exploit the increased availability of a large number of potential predictors. Consider the problem of forecasting $y_{t+h|t}$ based on its own current and past values as well as those of N additional variables, x_t. Of particular interest is the case in which the number of predictors, N, exceeds the number of time series observations, T. In that case, principal components analysis provides a convenient way of extracting a low-dimensional vector of common factors from the original data-set x_t (Stock and Watson 2002a, b). Forecasts that incorporate estimated common factors have proved successful in many cases in reducing forecast errors relative to traditional time series forecasting methods. Boivin and Ng (2005) provide a systematic comparison of alternative *factor model forecasts*. Another promising forecasting method is *Bayesian model averaging* across alternative forecasting models (Raftery et al. 1997). The latter method builds on the literature on forecast combinations (Bates and Granger 1969).

Dynamic Econometric Modelling

There is a close connection between multivariate time-series models and the structural, reduced and final forms of dynamic econometric models; the standard simultaneous-equations model (SEM) is a specific and restricted case.

Suppose that a vector of observed variables y_t may be subdivided into two classes of variables, 'exogenous', $\{x_t\}$, and endogenous, $\{z_t\}$. A dynamic, multivariate simultaneous linear system may be written.

$$\begin{bmatrix} \Psi_{11}(L) & \Psi_{12}(L) \\ 0 & \Psi_{22}(L) \end{bmatrix} \begin{pmatrix} z_t \\ x_t \end{pmatrix} = \begin{bmatrix} \Theta_{11}(L) & 0 \\ 0 & \Theta_{22}(L) \end{bmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \tag{41}$$

where $\Psi_{ij}(\cdot)$ and $\Theta_{ij}(\cdot)$, $i, j = 1, 2$ are matrix polynomials in the lag operator L. Such systems are known as vector ARMAX models and conditions for their identification are given by Hatanaka (1975). The reduced form of the system is

obtained by expressing zt as a function of lagged endogenous and current and lagged exogenous variables. The final form is then obtained by eliminating the lagged endogenous variables (see Zellner and Palm 1974; Wallis 1977).

Structural Vector Autoregressions

An important special case of the dynamic SEM is the structural vector autoregressive model in which all variables are presumed endogenous, the lag structure is unrestricted up to some order p, and identification of the structural from is achieved by imposing restrictions on the correlation structure of the structural innovations (Sims 1980). The most common form of the structural VAR(p) model imposes restrictions on the contemporaneous interaction of structural innovations. Consider the structural form for a K-dimensional vector $\{xt\}$, $t = 1, \ldots, T$:

$$B_0 x_t = \sum_{i=1}^{p} B_i x_{t-i} + \eta_t, \qquad (42)$$

where $\eta_t \sim (0, \Sigma_\eta)$ denotes the $(K \times 1)$ vector of serially uncorrelated structural innovations (or shocks) and Bi, $i = 0, \ldots, p$, the $(K \times K)$ coefficient matrices. Without loss of generality, let $\Sigma\eta = I$. The corresponding reduced form is

$$x_t = \sum_{i=1}^{p} B_0^{-1} B_i x_t + B_0^{-1} \eta_t = \sum_{i=1}^{p} A_i x_t + \varepsilon_t \quad (43)$$

where $\varepsilon_t \sim (0, \Sigma_\varepsilon)$ Since $\varepsilon_t = B_0^{-1}\eta_t$, it follows that $\sum_\varepsilon = B_0^{-1} B_0^{-1}\prime$. Given a consistent estimate of the reduced form parameters Ai, $i = 1, \ldots, p$, and Σ_ε, the elements of B_0^{-1} will be exactly identified after imposing $K(K-1)/2$ restrictions on the parameters of B_0^{-1} that reflect the presumed structure of the economy. Given estimates of B_0^{-1} and A_i, $i = 1, \ldots, p$, estimates of the remaining structural parameters may be recovered from $B_i = B_0 A_i$.

In practice, the number of restrictions that can be economically motivated may be smaller or larger than $K(K-1)/2$. Alternative estimation strategies that remain valid in the over-identified case include the generalized method of moments (Bernanke 1986) and maximum likelihood (Sims 1986). An instrumental variable interpretation of VAR estimation is discussed in Shapiro and Watson (1988). Semi-structural VAR models that are only partially identified have been proposed by Bernanke and Mihov (1998).

Alternative identification strategies may involve putting restrictions on the long-run behaviour of economic variables (Blanchard and Quah 1989; King et al. 1991) or on the sign and/or shape of the impulse responses (Faust 1998). Other possibilities include identification via heteroskedasticity (Rigobon 2003) or the use of high-frequency data (Faust et al. 2004).

The estimates of the structural VAR form may be used to compute the dynamic responses of the endogenous variables to a given structural shock, variance decompositions that measure the average contribution of each structural shock to the overall variability of the data, and historical decompositions of the path of x_t based on the contribution of each structural shock.

Conclusions

The literature on time series analysis has made considerable strides since the 1980s. The advances have been conceptual, theoretical and methodological. The increased availability of inexpensive personal computers in particular has revolutionized the implementation of time series techniques by shifting the emphasis from closed-form analytic solutions towards numerical and simulation methods. The ongoing improvements in information technology, broadly defined to include not only processing speed but also data collection and storage capabilities, are likely to transform the field even further. For example, the increased availability of large cross-sections of time series data, the introduction of ultra high-frequency data, the electronic collection of micro-level time series data (such as web-based data or scanner data), and the increased availability of data in real time all are creating new applications and spurring interest in the development of new methods of time series analysis. These developments already have brought together the fields of empirical finance and time series econometrics, resulting in the emergence of the new and fertile field of financial

econometrics. As the use of time series methods becomes more widespread in applied fields, there will be increasing interest in the development of methods that can be adapted to the specific objectives of the end user. Another question of growing importance is how to deal with rapidly evolving economic environments in the form of structural breaks and other model instabilities. Finally, the improvement of structural time series models for macroeconomic policy analysis will remain a central task if time series analysis is to retain its importance for economic policymaking.

See Also

- ▶ Cointegration
- ▶ Factor Models
- ▶ Forecasting
- ▶ Long Memory Models
- ▶ Maximum Likelihood
- ▶ Non-linear Time Series Analysis
- ▶ Seasonal Adjustment
- ▶ Spectral Analysis
- ▶ Spurious Regressions
- ▶ Trend/Cycle Decomposition
- ▶ Vector Autoregressions

Bibliography

Adelman, I. 1965. Long cycles – Fact or artifact? *American Economic Review* 60: 443–463.

Akaike, H. 1970. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22: 203–217.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

Andersen, T., T. Bollerslev, P. Christoffersen, and F. Diebold. 2006a. Volatility and correlation forecasting. In *Handbook of economic forecasting*, ed. G. Elliott, C. Granger, and A. Zimmermann. Amsterdam: North-Holland.

Andersen, T., T. Bollerslev, and F. Diebold. 2006b. Parametric and nonparametric volatility measurement. In *Handbook of financial economics*, ed. L. Hansen and Y. Ait-Sahalia. Amsterdam: North-Holland.

Anderson, T. 1971. *The statistical analysis of time series*. New York: Wiley.

Anderson, T. 1977. Estimation for autoregressive moving average models in the time and frequency domains. *Annals of Statistics* 5: 842–865.

Anderson, T., and A. Takemura. 1986. Why do noninvertible moving averages occur? *Journal of Time Series Analysis* 7: 235–254.

Ansley, C. 1979. An algorithm for the exact likelihood of a mixed autoregressivemoving average process. *Biometrika* 66: 59–65.

Baillie, R. 1996. Long memory processes and fractional integration in econometrics. *Journal of Econometrics* 73: 5–59.

Baillie, R., T. Bollerslev, and H.-O. Mikkelsen. 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 74: 3–30.

Bates, J., and C. Granger. 1969. The combination of forecasts. *Operational Research Quarterly* 20: 451–468.

Bell, W., and S. Hillmer. 1984. Issues involved with seasonal analysis of economic time series. *Journal of Business and Economic Statistics* 2: 291–349.

Bernanke, B. 1986. Alternative explanations of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy* 25: 49–100.

Bernanke, B., and I. Mihov. 1998. Measuring monetary policy. *Quarterly Journal of Economics* 113: 869–902.

Beveridge, W. 1921. Weather and harvest cycles. *Economic Journal* 31: 429–452.

Beveridge, W. 1922. Wheat prices and rainfall in western Europe. *Journal of the Royal Statistical Society* 85: 412–459.

Blanchard, O., and D. Quah. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79: 655–673.

Boivin, J., and S. Ng. 2005. Understanding and comparing factor-based forecasts. *International Journal of Central Banking* 1 (3): 117–152.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–327.

Box, G., and G. Jenkins. 1970. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Box, G., S. Hillmer, and G. Tiao. 1978. Analysis and modeling of seasonal time series. In *Seasonal analysis of economic time series*, ed. A. Zellner. Washington, DC: Bureau of the Census, Department of Commerce.

Brillinger, D. 1975. *Time series: Data analysis and theory*. New York: Holt.

Burman, J. 1980. Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society. Series A* 143: 321–337.

Burridge, P., and K. Wallis. 1985. Calculating the variance of seasonally adjusted series. *Journal of the American Statistical Association* 80: 541–552.

Cappé, O., E. Moulines, and T. Ryden. 2005. *Inference in hidden Markov models*. New York: Springer-Verlag.

Chambers, M. 1998. Long memory and aggregation in macroeconomic time series. *International Economic Review* 39: 1053–1072.

Christiano, L., and T. Fitzgerald. 2003. The band pass filter. *International Economic Review* 44: 435–465.

Christoffersen, P., and F. Diebold. 1996. Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics* 11: 561–571.

Christoffersen, P., and F. Diebold. 1997. Optimal prediction under asymmetric loss. *Econometric Theory* 13: 808–817.

Cournot, A. 1838. *Researches into the mathematical principles of the theory of wealth*. Trans. N. Bacon. New York: Macmillan, 1927.

Diebold, F., and A. Inoue. 2001. Long memory and regime switching. *Journal of Econometrics* 105: 131–159.

Doob, J. 1953. *Stochastic processes*. New York: Wiley.

Durbin, J., and S. Koopman. 2001. *Time series analysis by state space methods*. Oxford: Oxford University Press.

Durlauf, S., and P. Phillips. 1988. Trends versus random walks in time series analysis. *Econometrica* 56: 1333–1354.

Elliott, G. 1998. The robustness of co-integration methods when regressors almost have unit roots. *Econometrica* 66: 49–58.

Engle, R. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* 50: 987–1008.

Engle, R., and C. Granger. 1987. Co-integration and error correction: Representation, estimation and testing. *Econometrica* 55: 251–276.

Engle, R., and C. Granger. 1991. *Long run economic relations: Readings in co-integration*. Oxford: Oxford University Press.

Faust, J. 1998. The robustness of identified VAR conclusions about money. *Carnegie-Rochester Conference Series on Public Policy* 49: 207–244.

Faust, J., E. Swanson, and J. Wright. 2004. Identifying VARs based on high frequency futures data. *Journal of Monetary Economics* 51: 1107–1131.

Fishman, G. 1969. *Spectral methods in econometrics*. Cambridge: Harvard University Press.

Franses, P. 1996. *Periodicity and stochastic trends in economic time series*. Oxford: Oxford University Press.

Fuller, W. 1976. *Introduction to statistical time series*. New York: Wiley.

Geweke, J., and S. Porter-Hudak. 1983. The estimation and application of long memory series models. *Journal of Time Series Analysis* 4: 221–238.

Ghysels, E., and D. Osborn. 2001. *The econometric analysis of seasonal time series*. Cambridge: Cambridge University Press.

Granger, C. 1966. The typical spectral shape of an economic variable. *Econometrica* 34: 150–161.

Granger, C. 1969. Prediction with a generalized cost of error function. *Operations Research Quarterly* 20: 199–207.

Granger, C. 1980. Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* 14: 227–238.

Granger, C. 1981. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16: 121–130.

Granger, C., and P. Newbold. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2: 111–120.

Granger, C., and P. Newbold. 1977. *Forecasting economic time series*. New York: Academic Press.

Granger, C., and T. Teräsvirta. 1993. *Modelling nonlinear economic relationships*. Oxford: Oxford University Press.

Grether, D., and M. Nerlove. 1970. Some properties of 'optimal' seasonal adjustment. *Econometrica* 38: 682–703.

Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384.

Hannan, E. 1960. *Time series analysis*. London: Methuen.

Hannan, E. 1969. The estimation of mixed moving average autoregressive systems. *Biometrika* 56: 223–225.

Hannan, E. 1970. *Multiple time series*. New York: Wiley.

Hannan, E. 1971. The identification problem for multiple equation systems with moving average errors. *Econometrica* 39: 751–765.

Hannan, E. 1976. The identification and parameterization of ARMAX and state space forms. *Econometrica* 44: 713–723.

Hannan, E. 1979. The statistical theory of linear systems. In *Developments in statistics*, ed. P. Krishnaiah. New York: Academic Press.

Harvey, A. 1981. *Time series models*. Oxford: Allan.

Harvey, A. 1984. A unified view of statistical forecasting procedures. *Journal of Forecasting* 3: 245–275.

Harvey, A. 1989. *Forecasting. Structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

Harvey, A., and S. Peters. 1990. Estimation procedures for structural time series models. *Journal of Forecasting* 9: 89–108.

Harvey, A., and P. Todd. 1984. Forecasting economic time series with structural and Box–Jenkins models: A case study (with discussion). *Journal of Business and Economic Statistics* 1: 299–315.

Hatanaka, M. 1975. On the global identification of the dynamic simultaneous equations model with stationary disturbances. *International Economic Review* 16: 545–554.

Hillmer, S., and G. Tiao. 1982. An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* 77: 63–70.

Hillmer, S., W. Bell, and G. Tiao. 1983. Modeling considerations in the seasonal analysis of economic time series. In *Applied time series analysis of economic data*, ed. A. Zellner. Washington, DC: Bureau of the Census, Department of Commerce.

Hodrick, R., and E. Prescott. 1997. Postwar US business cycles: An empirical investigation. *Journal of Money, Credit and Banking* 29: 1–16.

Holt, C. 1957. *Forecasting seasonals and trends by exponentially weighted moving averages*. ONR Research Memorandum No. 52, Carnegie Institute of Technology.

Hylleberg, S. 1992. *Modelling seasonality*. Oxford: Oxford University Press.

Jevons, W. 1884. *Investigations in currency and finance*. London: Macmillan.

Johansen, S. 1995. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford: Oxford University Press.

Kalman, R. 1960. A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers – Journal of Basic Engineering* 82 (Series D): 35–45.

King, R., C. Plosser, J. Stock, and M. Watson. 1991. Stochastic trends and economic fluctuations. *American Economic Review* 81: 819–840.

Kolmogorov, A. 1941. Interpolation und Extrapolation von stationären zufälligen Folgen. *Bulletin of the Academy Science (Nauk), USSR, Mathematical Series* 5: 3–14.

Koopmans, L. 1974. *The spectral analysis of time series*. New York: Academic Press.

Kuznets, S. 1961. *Capital and the American economy: Its formation and financing*. New York: Princeton University Press for the National Bureau of Economic Research.

Maravall, A. 1981. *Desestacionalization y Politica Monetaria*. Economic studies no. 19. Madrid: Bank of Spain.

Maravall, A. 1984. *Model-based treatment of a manic depressive series*. Working paper, Bank of Spain.

Muth, J. 1960. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55: 299–305.

Nelson, C., and H. Kang. 1981. Spurious periodicity in inappropriately detrended time series. *Econometrica* 49: 741–752.

Nelson, C., and C. Plosser. 1982. Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics* 10: 139–162.

Nerlove, M. 1964. Spectral analysis of seasonal adjustment procedures. *Econometrica* 32: 241–286.

Nerlove, M. 1965. A comparison of a modified Hannan and the BLS seasonal adjustment filters. *Journal of the American Statistical Association* 60: 442–491.

Nerlove, M. 1967. Distributed lags and unobserved components in economic time nseries. In *Ten economic essays in the tradition of Irving Fisher*, ed. W. Fellner et al. New York: Wiley.

Nerlove, M., and S. Wage. 1964. On the optimality of adaptive forecasting. *Management Science* 10: 207–224.

Nerlove, M., D. Grether, and J. Carvalho. 1979. *Analysis of economic time series*. New York: Academic Press.

Newbold, P. 1974. The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika* 61: 423–426.

Parzen, E. 1961. An approach to time series analysis. *Annals of Mathematical Statistics* 32: 951–989.

Phillips, P. 1986. Understanding spurious regressions in econometrics. *Journal of Econometrics* 33: 311–340.

Phillips, P., and S. Durlauf. 1986. Multiple time series regression with integrated processes. *Review of Economic Studies* 53: 473–495.

Phillips, P., and B. Hansen. 1990. Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies* 57: 99–125.

Pierce, D. 1978. Seasonal adjustment when both deterministic and stochastic seasonality are present. In *Seasonal analysis of economic time series*, ed. A. Zellner. Washington, DC: Bureau of the Census, Department of Commerce.

Pierce, D. 1979. Signal extraction error in nonstationary time series. *Annals of Statistics* 7: 1303–1320.

Priestley, M. 1981. *Spectral analysis and time series*. New York: Academic Press.

Raftery, A., D. Madigan, and J. Hoeting. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92: 179–191.

Ravn, M., and H. Uhlig. 2002. On adjusting the HP-filter for the frequency of observations. *Review of Economics and Statistics* 84: 371–376.

Rigobon, R. 2003. Identification through heteroskedasticity. *Review of Economics and Statistics* 85: 777–792.

Rudebusch, G. 1993. The uncertain unit root in real GNP. *American Economic Review* 83: 264–272.

Sargan, J., and A. Bhargava. 1983. Maximum likelihood estimation of regression models with moving average errors when the root lies on the unit circle. *Econometrica* 51: 799–820.

Schuster, A. 1898. On the investigation of hidden periodicities with application to the supposed 26-day period of meteorological phenomena. *Terrestrial Magnetism and Atmospheric Electricity* [now *Journal of Geophysical Research*] 3: 13–41.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Schweppe, F. 1965. Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory* 11: 61–70.

Shapiro, M., and M. Watson. 1988. Sources of business cycle fluctuations. *NBER Macroeconomics Annual* 3: 111–156.

Shephard, N. 2005. *Stochastic volatility: Selected readings*. Oxford: Oxford University Press.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.

Sims, C. 1986. Are forecasting models usable for policy analysis? *Quarterly Review, Federal Reserve Bank of Minneapolis* 10: 2–16.

Sims, C., J. Stock, and M. Watson. 1991. Inference in linear time series models with some unit roots. *Econometrica* 58: 113–144.

Slutsky, E. 1927. The summation of random causes as the source of cyclic processes. *Econometrica* 5 (April 1937): 105–46.

Sowell, F. 1992. Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics* 53: 165–188.

Stock, J. 1991. Confidence intervals for the largest autoregressive root in US economic time series. *Journal of Monetary Economics* 28: 435–460.
Stock, J., and M. Watson. 1993. A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* 61: 783–820.
Stock, J., and M. Watson. 2002a. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167–1179.
Stock, J., and M. Watson. 2002b. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20: 147–162.
Stokes, G. 1879. Note on searching for hidden periodicities. *Proceedings of the Royal Society* 29: 122–125.
Theil, H., and S. Wage. 1964. Some observations on adaptive forecasting. *Management Science* 10: 198–206.
Wallis, K. 1977. Multiple time series analysis and the final form of econometric models. *Econometrica* 45: 1481–1497.
Whittle, P. 1963. *Prediction and regulation by linear least-squares methods*. London: English Universities Press.
Wiener, N. 1949. *The extrapolation, interpolation and smoothing of stationary time series with engineering applications*. New York: Wiley.
Winters, P. 1960. Forecasting sales by exponentially weighted moving averages. *Management Science* 6: 324–342.
Wold, H. 1938. *A study in the analysis of stationary time series*. Stockholm: Almqvist and Wiksell.
Yule, G. 1921. On the time-correlation problem, with special reference to the variate difference correlation method. *Journal of the Royal Statistical Society* 84: 497–526.
Yule, G. 1926. Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89: 1–64.
Yule, G. 1927. On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London A* 226: 267–298.
Zellner, A., and F. Palm. 1974. Time series analysis and simultaneous equation econometric models. *Journal of Econometrics* 2: 17–54.

Time Use

Duncan Ironmonger

Abstract

Time is a finite, irreplaceable resource, and unlike money is equally distributed. Statistics of time use and money use can be combined, giving two-dimensional pictures of individual economic behaviour and the national economy. A framework for the *microeconomics* of time use and household production was established in the 1960s. In coming decades, a *macroeconomics* of time use and household production will arise. Within a production, consumption and investment framework, this will employ continuous national time accounts and satellite accounts of the household economy. Consequently the household's true economic role and its powerful interactions with the market will be revealed.

Keywords

American Time Use Survey (ATUS); Bureau of Labor Statistics (BLS); Child care; Consumption activities; Diary-based surveys of time use; Education; Gross Economic Product (GEP); Gross Household Product (GHP); Gross Market Product (GMP); Household economy; Household production; Household productivity; Household technology; Human capital activities; Investment in human capital; Labour statistics; Leisure; Macroeconomic modelling of the household economy; Microeconomics of household production and consumption; Multinational Time Use Study (MTUS); National time accounts; Paid work; Production activities; Reid, M.; Robbins, L.; Satellite accounts of the household economy; Social interaction and recreation; System of National Accounts (SNA); Third-person test; Time income and expenditure; Time Resources and Time Use Account; Time use; Work-leisure trade-off

JEL Classifications

D11

Time is a finite, irreplaceable resource available to every man, woman and child in equal amounts of 168 hours per week over the course of life. Time use refers to the allocation of time to alternative uses such as sleep, leisure or work.

Time is perhaps the most fundamental scarce resource; unlike money income or wealth, it is equally distributed; and how well or wastefully it is used largely determines the progress, achievement and well-being of individuals, families, communities and societies.

One way to assess progress and well-being is to measure aggregate changes in the uses of time rather than, or in addition to, the usual monetary statistics of national income and expenditure. National time accounts are more comprehensive than money accounts as they simultaneously measure the productive time spent in both the market and the household, as well as the time spent in consumption of outputs from both.

The advent of radio astronomy in 1957, using a wider spectrum than that provided by the visible light frequencies, opened up new views of the astronomical universe. Similarly, official statistical organizations are beginning to provide new views of the economic and social universe by observing the world through a 'time frequency' rather than the more visible 'money frequency'.

Researching Time Use and Human Behaviour

Time use economics is the scientific study of human behaviour concerning the allocation of time between alternative uses (cf. Robbins 1932, p. 16).

Time use can be usefully considered under the three comprehensive categories of production, consumption and investment. These accord with the theoretical framework of macroeconomics. Time use data show how individuals and households allocate their time between production (paid work and unpaid work), consumption (meals, television, social interaction and recreation) and human capital formation and maintenance (education, self-care and sleep).

Margaret Reid provided the well-known 'third person' test for separating production from consumption: 'If an activity is of such character that it might be delegated to a paid worker, then that activity shall be deemed productive' (Reid 1934, p. 11).

To facilitate analysis, macroeconomics makes a further distinction between resources for immediate use – consumption – and those for use over a longer time period – investment. Accordingly, time spent learning a skill or gaining knowledge in education is clearly investment in human capital. Similarly, time spent in sleep and self-care can be regarded as a necessary daily investment to maintain functioning minds and bodies.

When people spend time, in economic terms they are effectively allocating time between market production, household production, consumption and investment. Modelling of household time-allocation decisions goes beyond understanding the simple work-leisure trade-off. It provides knowledge of the detailed interactions between production, consumption and investment activities. It also provides a framework for the analysis of the derived demands for market commodities implicit in household production and consumption.

In a much-quoted article in *Scientific American,* Vanek (1974) surveys the time spent in household production in the United States over a period of 40 years. Juster and Stafford (1991) provides a comprehensive appraisal of the importance of time allocation as an analytic construct and a review of what had been learned from time allocation data in modelling economic behaviour and the dynamics of economic change. Robinson and Godbey (1999) give an account of the American data from 1965, 1975, 1985 and 1995, and Gershuny (2000) comprehensively surveys the time use data for Britain.

Becker (1965) offers a theoretical framework for the microeconomics of time use by including the cost of time on the same footing as the cost of market goods and by assuming households 'combine time and market goods to produce more basic commodities that directly enter their utility functions' (1965, p. 494). This neoclassical approach to a theory has been criticized on many grounds, summarized recently by Folbre (2004).

While neoclassical micro-theory provides a useful framework for a *microeconomics of household production and consumption*, it completely disregards the full macroeconomic magnitude of household production. In most

developed countries more labour is required for household production than for market production (Goldschmidt-Clermont and Pagnossin-Aligisakis 1995). Indeed, if the unpaid time spent in caring for children is fully measured, including the 'parallel' or 'secondary' time when other activities are simultaneously undertaken, total work in household production is considerably greater than market work (Ironmonger 2004).

Existing official labour statistics are misleading indicators of total work as they ignore labour inputs to household production. They measure only that time spent on activities within the production boundary of the System of National Accounts (SNA). A broader definition of work includes unpaid work outside the SNA boundary but within the general production boundary. This non-SNA work is employment in household production – the provision of meals, clean clothes, accommodation, care and transport by households for themselves or other households without remuneration (Ironmonger 2001).

Time use economics is much more than the microeconomics of choice. It includes the study of the large non-monetary economy where households combine time, *their own capital* and intermediate inputs from the market to produce services that compete with market services. Including the use of household capital, the Gross Household Product (GHP) of the household economy is comparable to the Gross Market Product (GMP) of the market (Ironmonger 1996a).

Time use data will provide the raw materials for a new *macroeconomics of the household economy.* A principal objective of the recent Eurostat and United States initiatives to collect more time use data has been not only to value unpaid work but to produce satellite accounts giving monetary valuations of the household economy (Varjonen et al. 1999). To be effective as a new set of tools for macroeconomic modelling of the household economy, these accounts and the national time accounts will need to be a continuous time series of annual or even quarterly data covering a run of years.

Considering its enormous magnitude, the performance and rate of growth of the household economy deserve as much attention as that given to the market economy. Macroeconomic modelling of the household economy could help explain the rate of change of household production, the impact of changes in household technology on productivity, and the effects of monetary and fiscal policies on GHP and its broad components.

Modelling of the total economic system – especially the dynamic interactions between the market and household economies – will be greatly improved by using a continuous record of time use taking account of changing economic, social and technological factors. Booms and slumps affect the mix between paid and unpaid work in the market and household economies; new products and technologies can quickly affect the relative productivity of these sectors, and tax policies can alter the relative attractiveness of paid work, unpaid work and leisure.

Using national time accounts and satellite accounts of household production, a macroeconomic model of the household economy could be constructed and linked with a macroeconomic model of the market (Ironmonger 1994, 1996b). Such a comprehensive model could investigate the cyclical and long-run relationships between market and household production. One hypothesis worth verification is that the two spheres of production vary in a counter-cyclical pattern (Ironmonger 1989). If this is so, the amplitude of the fluctuation of Gross Economic Product (GEP, the sum of GHP and GMP) will be less than either component.

Measuring Time Use

The accurate scientific measurement of how people use time began with independent surveys in a number of countries, particularly in the USSR and the United States in the 1920s. A major advance was made in 1965 when Alexander Szalai directed internationally comparable, diary-based surveys in 12 countries under the sponsorship of UNESCO and the International Social Science Council (Szalai 1972). Subsequent official measurement of time use included Norway at ten-year

Time Use, Table 1 National time accounts: time resources and time use account, United States of America, 2003

	Women	Men	Children	Women	Men	Children	Total
Population (million)				118.1	111.9	60.7	290.7
	Average hours per week			Million hours per week			
Time resources Total time income	**168.0**	**168.0**	**168.0**	**19,841**	**18,799**	**10,198**	**48,838**
Time use Time expenditure							
Production activities							
Household economy (non-SNA production)	32.6	20.3	*5.0*	3,850	2,272	*304*	6,425[a]
Market economy (SNA production)	20.1	32.0	–	2,374	3,581	–	5,955
Total production activity	**52.7**	**52.3**	*5.0*	**6,224**	**5,952**	*304*	**12,380**
Consumption activities							
Eating and drinking	8.3	8.7	*10.0*	980	974	*607*	2,561
Watching TV	16.8	19.3	*26.0*	1,984	2,160	*1,578*	5,722
Social and recreation	16.9	18.6	*24.0*	1,996	2,081	*1,457*	5,534
Total consumption activity	**42.0**	**46.6**	*60.0*	**4,960**	**5,215**	*3,642*	**13,817**
Human capital activities							
Education	3.5	3.2	*22.0*	413	358	*1,335*	2,107
Self-care	6.2	4.6	*5.0*	732	515	*304*	1,550
Sleep	60.6	59.3	*74.0*	7,157	6,636	*4,492*	18,284
Total human capital activity	**70.3**	**67.1**	*101.0*	**8,302**	**7,508**	*6,131*	**21,942**
Telephone, etc.	3.0	2.0	*2.0*	354	224	*121*	700
Total time expenditure	**168.0**	**168.0**	**168.0**	**19,841**	**18,799**	**10,198**	**48,838**

Source: Estimates of the Households Research Unit, Department of Economics, University of Melbourne based on Bureau of Labor Statistics, Time-Use Survey – First Results Announced by BLS, 14 September 2004, http://www.bls.gov/tus/home.htm and Population Division, US Census Bureau

[a]ATUS provides estimates of secondary child care, but restricted to times when respondents are awake and children are not in bed. If these estimates are used, this figure would increase by about 19 per cent to 7,600 million hours per week if, without double-counting other household work, the hours spent in caring for children younger than 13 years as a secondary activity is included. Most of his extra time occurred simultaneously with consumption activity but some overlapped with market work, education and self-care

intervals from 1971, the Netherlands at five-year intervals from 1975, Canada in 1981, 1986, 1992 and 1998, and Australia in 1987, 1992 and 1997. Several developing countries – for example, India in 1998–9 and South Africa in 2001 – have now conducted official time use surveys.

This development in scientific measurement culminated in the harmonized surveys across some 20 European countries in 1998–2003. They used what is regarded as the most valid technique to measure time use – a diary recording the chronology of various time uses over one or more days from a representative sample of the population. The European surveys collected one weekday and one weekend day from people aged ten upwards; the Indian surveyed one day from age six upwards. Although the 1965 surveys were in one city in one month of the year, the subsequent official surveys cover the whole population, urban and rural, and all seasons of the year.

Research on time use by universities, government agencies and private business has been greatly facilitated by access to the unit record files of individual diary days. The Multinational Time Use Study (MTUS) contains a growing collection of these files covering 78 surveys in 27 countries, including the original 12 from the 1965 cross-national time budget study. The 'World' files of MTUS are a particularly rich resource for international comparison, as data

have been made as comparable as possible by providing common categories of time use and standard definitions of demographic and household variables.

In addition to the diary-based surveys of time use, some official statistical organizations make estimates, on a continuous basis, of time used in the market economy. For example, in Australia, since 1966, the monthly household survey of employment and unemployment also provides estimates of average hours of market work. In this survey the interviewer's question is: 'How many hours did you work last week in (all) your job(s)?' This is a 'stylized' question subject to biases that differ from the diary-based method. Each method – detailed diary or stylized question – has its own bias against 'reality'.

With the inauguration of the American Time Use Survey (ATUS) in January 2003, the Bureau of Labor Statistics and the US Census Bureau took a giant stride in the scientific measurement of time use. In the world's first continuous survey, diary data on time use are collected each month from a representative sample of adults aged 15 or more years.

This groundbreaking response to fill the biggest single gap in the Federal Statistical System of the United States was 12 years in development. It traces its origin to 1991 when a bill introduced in the 102nd Congress called for the Bureau of Labor Statistics to 'conduct time-use surveys of unremunerated work performed in the United States and to calculate the monetary value of such work' (Horrigan and Herz 2004).

Table 1 shows the Time Resources and Time Use Account for the United States of America for 2003 derived from the first results of ATUS and arranged according to the macroeconomic categories of production, consumption and investment. Estimates of the time use of children under 15 based on other surveys have been included to complete the table.

As the way children spend time is a critical issue for many research and policy purposes, time use data should be extended to include children. An innovative feature of the new longitudinal study of Australian children is a time-use diary where parents record details of what their child did in two 24-hour periods (Australian Institute of Family Studies 2005).

Conclusion

The development of national time accounts and satellite accounts of the household economy should be an interactive process between researchers, model builders, policymakers and official statisticians. The national money income and expenditure accounts developed this way.

At the beginning of the 21st century statistical measurements are starting to provide the raw materials for a *macroeconomics of the household economy.* New continuous time use observations will provide regular national time accounts and satellite accounts of household production.

The new data will enable the building of more relevant economy-wide models to explore the continuing interactions between the household and the formal market and public sectors of the economic system. The true macroeconomic role of the household will become clear, not only as the place for consumption and leisure but as the largest user of labour time in economic production. The new household-based models should provide better understanding of issues of work and leisure, and produce policies conducive to a more equitable and satisfactory allocation of time.

See Also

- ▶ Becker, Gary S. (Born 1930)
- ▶ Childcare
- ▶ Household Production and Public Goods
- ▶ Household Surveys
- ▶ Human Capital
- ▶ Intrahousehold Welfare
- ▶ Leisure
- ▶ National Accounting, History of
- ▶ Robbins, Lionel Charles (1898–1984)
- ▶ Value of Time
- ▶ Women's Work and Wages

Bibliography

Australian Institute of Family Studies. 2005. *Growing up in Australia: The longitudinal study of Australian children: 2004 annual report*. Canberra: AIFS.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Folbre, N. 2004. A theory of the misallocation of time. In *Family time: The social organization of care*, ed. N. Folbre and M. Bittman. London: Routledge.

Gershuny, J. 2000. *Changing times*. Oxford: Oxford University Press.

Goldschmidt-Clermont, L., and E. Pagnossin-Aligisakis. 1995. Measures of unrecorded economic activities in fourteen countries. Occasional Paper No. 20. New York: Human Development Report Office, UNDP.

Horrigan, M., and D. Herz. 2004. Planning, designing, and executing the BLS American time-use survey. *Monthly Labor Review* 127(10): 3–19.

Ironmonger, D. 1989. Households and the household economy. In *Households work: Productive activities, women and income in the household economy*, ed. D. Ironmonger. Sydney: Allen and Unwin.

Ironmonger, D. 1994. Modeling the household economy. In *Economics, econometrics and the LINK: Essays in honor of Lawrence R. Klein*, ed. M. Dutta. Amsterdam: North Holland.

Ironmonger, D. 1996a. Counting outputs, capital inputs and caring labor: Estimating gross household product. *Feminist Economics* 2: 37–64.

Ironmonger, D. 1996b. Time use and satellite accounts for modelling the household economy. Paper presented at the IARIW 24th general conference, Lillehammer, August.

Ironmonger, D. 2001. Household production. In *International encyclopedia of the social & behavioral sciences*, vol. 10, ed. N. Smelser and P. Baltes. Oxford: Elsevier Science.

Ironmonger, D. 2004. Bringing up Bobby and Betty: The inputs and outputs of childcare time. In *Family time: The social organization of care*, ed. N. Folbre and M. Bittman. London: Routledge.

Juster, F., and F. Stafford. 1991. The allocation of time: Empirical findings, behavioural models, and problems of measurement. *Journal of Economic Literature* 29: 471–522.

Reid, M. 1934. *Economics of household production*. New York: Wiley.

Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.

Robinson, J., and G. Godbey. 1999. *Time for life: The surprising ways Americans spend their time*. University Park: Penn State Press.

Szalai, A. (ed.). 1972. *The uses of time: Daily activities of urban and suburban populations in twelve countries*. The Hague: Mouton.

Vanek, J. 1974. Time spent in housework. *Scientific American* 231: 116–120.

Varjonen, J., I. Niemi, E. Hamunen, H. Paakkonen, and T. Sandstrom. 1999. Proposal for a satellite account of household production. Working Paper No. 9/1999/A4/11. Brussels: Eurostat.

Timlin, Mabel Frances (1891–1976)

Robert W. Dimand

Keywords

American Economic Association; Canada, economics in; General equilibrium; Keynesianism; Timlin, M. F

JEL Classifications

B31

The Keynesian economist Mabel Timlin was the first tenured woman among Canadian economists, the first woman elected president of the Canadian Political Science Association (which then covered all social sciences, including economics), the first woman outside the natural sciences elected a Fellow of the Royal Society of Canada (1951), and one of the first ten women to serve on the executive committee of the American Economic Association (1958–60), despite becoming an assistant professor only in her 50th year, after a long career as an academic secretary. She was born in Forest Junction, Wisconsin, on 6 December 1891, and, after studying at the Milwaukee State Normal School, taught in Wisconsin and rural Saskatchewan. She became a secretary at the University of Saskatchewan in 1921, while studying for a BA there. At first Timlin intended to study economics, but after seeing the Department of Economics and Political Science at Saskatchewan she decided (probably correctly) that she could learn more economics on her own. She took a BA with great distinction in English in 1929, and then directed the university's

correspondence courses in economics. Mabel Timlin became an instructor in economics at the University of Saskatchewan in 1935, after completing graduate course work in economics at the University of Washington during summers and a six-month leave. Her doctoral dissertation at the University of Washington, supervised by the much younger Raymond Mikesell, was accepted in 1940 and published as *Keynesian Economics* (1942). In 1941, Timlin became an assistant professor of economics at the University of Saskatchewan (associate professor 1946, full professor 1950) and a member of the executive committee of the Canadian Political Science Association (Vice-President 1953–5, President 1959–60).

Keynesian Economics did more than introduce Keynesian theory into Canadian academic life. Timlin offered one of the early general equilibrium interpretations of John Maynard Keynes's *General Theory*, and was particularly noteworthy in treating it as a system of shifting equilibrium, presented with innovative diagrams on which she collaborated with the eminent geometer H.S.M. Coxeter. Timlin began work on *Keynesian Economics* in 1935, before Keynes published his *General Theory*: Benjamin Higgins had come to Saskatoon from the London School of Economics in 1935 for a one-year appointment, carrying a copy of the summary of Keynes's Cambridge lectures that Robert Bryce had presented in Friedrich Hayek's LSE seminar.

Beyond her work on Keynes, Timlin also expounded international developments in welfare economics and general equilibrium analysis to a Canadian audience more used to historical and institutional economics than to formal theory (for example, Timlin, 1946). Timlin (1953) sharply criticized the Bank of Canada for failing to follow Keynesian countercyclical stabilization policies during the Korean War inflation. Much of her later work (for example, Timlin, 1951; 1958; 1960) concerned immigration policy, emphasizing the economic benefits of freer immigration.

Mabel Timlin never married. Generations of former students were her extended family. She remained active as a scholar long after her official retirement in 1959, publishing a major report on the social sciences in Canada in 1968. She remained devoted to the University of Saskatchewan despite job offers from such institutions as the University of Toronto, and died in Saskatoon on 19 September 1976.

See Also

- ▶ Canada, Economics in
- ▶ Keynesian Revolution
- ▶ Keynesianism

Selected Works

1942. *Keynesian economics.* Toronto: University of Toronto Press.

1946. General equilibrium analysis and public policy. *Canadian Journal of Economics and Political Science* 12: 483–495.

1947. John Maynard Keynes. *Canadian Journal of Economics and Political Science* 13: 363–365.

1951. *Does Canada need more people?* Toronto: Oxford University Press.

1953. Recent developments in Canadian monetary policy. *American Economic Review: Papers and Proceedings* 43: 42–53.

1955. Monetary stabilization policies and Keynesian theory. In *Post-Keynesian economics,* ed. K.R. Kurihara. London: George Allen & Unwin.

1958. Canadian immigration with special reference to the post-war period. In *International migration.* International Economic Association. London: Macmillan.

1960. Presidential address: Canada's immigration policy, 1896–1910. *Canadian Journal of Economics and Political Science* 26: 517–532.

1968. The social sciences in Canada: retrospect and prospect. In *The social sciences in Canada: Two studies,* ed. M.F. Timlin and A. Faucher. Ottawa: Social Science Research Council of Canada.

1977. (With biographical note by A.E. Safarian and introduction by L. Tarshis.) *Keynesian economics.* Toronto: McClelland and Stewart, Carleton Library.

Bibliography

Ainley, M.G. 1999. Mabel F. Timlin, 1891–1976: A woman economist in the world of men. *Atlantis: A Women's Studies Journal* 23: 28–38.

Spafford, S. 2000. *No ordinary academics: Economics and political science at the University of Saskatchewan, 1910–1960*. Toronto: University of Toronto Press.

Tinbergen, Jan (1903–1994)

Peter A. Cornelisse and Herman K. van Dijk

Abstract

Jan Tinbergen was the first Nobel Laureate in economics in 1969. This article presents a brief survey of his many contributions to economics, in particular to macroeconometric modelling, business cycle analysis, economic policymaking, development economics, income distribution, international economic integration and the optimal regime. It further emphasizes his desire to contribute to the solution of urgent socio-economic problems and his passion for a more humane world.

Keywords

Business cycles; Capability tax; Club of Rome; Cobweb model; Computable general equilibrium models; Development economics; Econometrics; Economic integration; Educational attainment; Foreign aid; Frisch, R.; Gravitation model; Great Depression; Haavelmo, T.; Haberler, G.; Harvard Barometer; Income distribution; Inequality; Inequality between nations; Innovation; International division of labour; International external effects; Keynes, J. M.; Lump sum taxes; Mitchell, W.; Models; Netherlands Bureau for Economic Policy Analysis; Optimal distribution of income; Persons, W.; Planning in stages; Positional-exchange criterion; Rent seeking; Slutzky, E.; Social welfare function; Statistical mechanics; Targets and instruments; Tax shifting; Taxation of income; Tinbergen rule; Tinbergen, J.; Velocity of integration; Yule, G

JEL Classifications

B31

Jan Tinbergen was born in The Hague, The Netherlands, on 12 April 1903, the first of five children in an intellectually stimulating family with a love of foreign languages. Eventually two of the children would win a Nobel Prize: Jan in economics (in 1969) and Niko, an ethologist, in physiology or medicine (in 1973).

Jan Tinbergen enrolled as a student of mathematical physics at Leiden University in 1921, where he obtained his doctorate in 1929. By that time he had already decided to switch to economics. From 1926 to 1928 Tinbergen worked as a conscientious objector to national military service, first in a convict prison and later, and of greater import to his subsequent career, at the Central Bureau of Statistics. He continued to work there until 1945. In 1933 he became extraordinary professor of statistics, mathematical economics and econometrics at the Netherlands School of Economics in Rotterdam. As a result of his quantitative approach to the study of economic dynamics, he was invited to the League of Nations in Geneva during the period 1936–8 in order to carry out statistical tests of business cycles theories. In 1945, at the end of the Second World War, Tinbergen was appointed as the first director of the Central Planning Bureau at The Hague. He held this position until 1955, when he became full professor of mathematical economics and development planning at the Netherlands School of Economics, later Erasmus University, Rotterdam. Throughout the 1960s and a part of the 1970s he acted as adviser to various international organizations and to governments of a considerable number of less-developed countries. He was elected chairman of the United Nations Committee on Development Planning in 1965 and held this position until 1972. In 1969 he was awarded, together with Ragnar Frisch, the first Nobel Prize in Economics. After his retirement as full professor in 1973 he held the Cleveringa Chair in Leiden

for two years. He continued to be involved in various research projects at old age. Jan Tinbergen died on 9 June 1994.

Personal Motivation

Already at an early age Tinbergen was profoundly impressed by the horrors of the Great War – subsequently numbered as the First World War – partly because of the fate of the Austrian refugee children his parents had lodged. Later, in Leiden as a student, when he was invited by his postman to join him on his rounds, he was appalled by the conditions of poverty in which the local population lived. Wishing to contribute to the struggle against such social evils, he decided to become an economist. This decision was characteristic of Tinbergen and his attitude towards economic science in his later life: his scientific contributions would always be inspired by the desire to tackle the social problems he observed. Paul Ehrenfest, professor of theoretical physics and Tinbergen's mentor in Leiden, was not unsympathetic towards the switch from physics to economics. Having made important contributions to statistical mechanics together with his wife Tatyana Afanasyeva, he called Tinbergen's attention to the possibilities that a mathematical representation of economic problems would offer. The dissertation on minimum problems in physics and economics that Tinbergen defended in 1929 bridged the two disciplines.

Econometric Modelling and Business Cycle Research

In 1969 Tinbergen was awarded, together with R. Frisch, the first Nobel Prize in Economics 'for having developed and applied dynamic models for the analysis of economic processes', as the Nobel Prize committee described it.

The desire to combat the socio-economic consequences of the Great Depression of the 1930s was Tinbergen's most important motivation for studying business cycles. In his inaugural address as extraordinary professor in 1933 he summarized his project as 'statistics and mathematics in the service of business cycle research'. His approach contrasted with previous approaches to business cycle research (for more details, see, for example, Morgan 1990 and Jolink 2003). After a 19th-century undertaking by Juglar (1862) ascribing the recurrent business crises in Europe and North America to credit crises, and Jevons's (1884) study pointing to agricultural production cycles connected with sunspot numbers, several research projects in the early 20th century were devoted to the construction of so-called business cycle barometers. The purpose was to measure economic fluctuations through a particular index (or set of indices) with the aim of giving warning signals for turning points that would lead to a depression. An example was the Harvard Index of Business Conditions, informally known as the Harvard Barometer, constructed by a team led by Persons (1919). Another well-known descriptive approach to the business cycle during this period had been initiated by Mitchell (1913). His work was followed by that of Yule (1927) and Slutzky (1927), who suggested that the cumulative effect of random shocks could be the cause of cyclical patterns in economic variables. Frisch (1933), co-recipient of the 1969 Nobel Prize, applied these ideas introducing econometric models in which impulse propagation mechanisms led to business cycles.

However useful it could be as a starting point, Tinbergen criticized descriptive analysis as being too vague for use in policy preparation, and started a quantitatively oriented research programme to explore the possible economic causes of the periodic upswings and downswings in economic activity. In an earlier theoretical study Aftalion (1927) had argued that lags in an economic model could generate cyclical variation in economic activity. Following up this argument, Tinbergen specified a first simple case using a system of difference equations to express lagged responses of supply to price changes in a market for a single good. He noted that the systematic fluctuations that could arise in such a system had been observed in an empirical study of the pork market by the German economist Hanau (1928), a phenomenon that became known as the 'cobweb

model' (Tinbergen 1979, presents additional relevant literature).

Tinbergen subsequently generalized the specification of dynamic equations with lagged adjustment processes to macroeconomic settings, arguing that fluctuations in components of national product, such as investment and consumption expenditures, would lead to business cycle fluctuations in general economic activity. In 1936 he published the first applied macroeconometric model (for the Netherlands). It was a dynamic model, consisting of 22 equations in 31 variables. Employing what we now see as basic statistical techniques like correlation and regression analysis, it was meant to be used for the analysis of the particularly pressing unemployment problem. The specification of consumption and employment in this model anticipated elements of Keynes's theory (1936). This modelling exercise resulted in a strong policy recommendation in favour of a devaluation of the Dutch guilder to tackle unemployment. But its importance for the economics profession was far more profound: for the first time the economic-policy debate had been based on empirically tested, quantitative economic analysis and not on rather informally stated economic theory, the so-called verbal approach. Thus, according to Solow (2004, p. 159), Tinbergen's work during this period 'was a major force in the transformation of economics from a discursive discipline into a model-building discipline'.

In 1936 Haberler had published a survey of theories on business cycles for the League of Nations. As a follow-up, and in reaction to the dynamic model for the Netherlands Tinbergen had published in that year, the same institution invited him to examine statistically which factors could be considered to contribute most to macroeconomic fluctuations. This project resulted in his two-volume book *Statistical Testing of Business Cycles Theories* (1939). The first volume contained a description of the methodology applied, while the second volume presented a dynamic macroeconometric model for the United States with the aim of studying business cycles in that country after the First World War. This model was not only considerably larger than the one for The Netherlands; as imports and exports were much less important for the United States, it also allowed a relatively undisturbed view of internal dynamic mechanisms. Subsequently, the US model was much refined and enlarged by Klein (1950) and Duesenberry et al. (1965). Tinbergen presented his views on the dynamics of business cycles and on objectives and instruments of business-cycle policy for a wider audience in Tinbergen (1943) and Tinbergen and Polak (1950).

Discussion with Keynes

The formulation of some relations in Tinbergen's 1936 model showed some resemblance to Keynes's theory. Nevertheless, in an article in the *Economic Journal* of 1939, Keynes was remarkably sceptical of Tinbergen's work. Keynes labeled Tinbergen's method of estimating the parameters of an econometric model and computing quantitative policy scenarios as 'statistical alchemy', arguing that this approach '. . .is a means of giving quantitative precision to what, in qualitative terms, we know already as the result of a complete theoretical analysis' (Keynes 1939, p. 560). Their widely diverging views on the relevance of quantitative economic analysis were also illustrated by Keynes's reaction to Tinbergen's estimate of the price elasticity of demand for exports. When, in 1919, Keynes had strongly criticized the excessive war indemnity payments enforced upon Germany after the First World War, his argument had depended critically on the value of this elasticity. Tinbergen empirically found this value to be minus 2, precisely the value that Keynes had assumed a priori in his study. When informed about this Keynes replied: 'How nice that you found the correct figure' (Kol and de Wolff 1993, p. 8).

Keynes's critical attitude towards macroeconometric modelling and analysis originated from his view that the underlying economic theory should be complete in the sense that it should include all relevant variables and set out in detail its causal and dynamic structure. Econometrics could be used only for measuring the relations ('curve fitting' was the term used); it could not refute economic hypotheses or evaluate economic models. Tinbergen, on the other hand, argued that

economic theories cannot be complete. Econometric research could be useful for scrutinizing elements of economic theories and for examining whether one theory describes reality better than another. Further, it could provide the numerical values of the coefficients in dynamic models that determine the cyclical and stability properties of a model, and, by applying a testing procedure of trial and error, it could yield suggestions for an improved specification of dynamic lags.

In this controversy Tinbergen's approach soon gained the upper hand as increasing numbers of economists, especially in the United States, noted its practical results in terms of model construction and verification, including forecasting. However, Keynes's comments on the role of expectations and uncertainty in macroeconometrics and on specification and simultaneous equation biases remained relevant. Haavelmo (1943) advocated the use of probability theory in bridging the gap between theory and data in business cycle analysis. Later these issues would become the subject of intensive debate and research.

Theory and Practice of Economic Policy

In 1945 Tinbergen was appointed director of the newly established Central Planning Bureau (CPB), an institution occupied with forecasting the effects of economic policy and advising the government on related matters (tasks which are more adequately captured by its present-day English name: Netherlands Bureau for Economic Policy Analysis). In the aftermath of the Second World War work at the CPB concentrated on the nation's pressing macroeconomic problems: a depleted capital stock, severe inflationary pressure, low levels of employment and an extreme shortage of foreign exchange.

In the economics discipline macroeconometric modelling had rapidly become accepted as a useful tool with the publication of such studies as by Klein (1950), Leontief (1950) and Klein and Goldberger (1955). But Tinbergen, having gained experience with the practice of policy preparation, felt the need for a systematic discussion of the logic of economic policy and of the use of models for policy purposes. It led to several monographs on the theory of economic policy (1952; 1956). He distinguished, among other things, between reforms (changes in the foundations of society), qualitative policy (changes in the structure of economic and social organization) and quantitative policy (changes in the instruments of economic policy). The latter could help to avoid the shortcomings of the traditional approach by offering a systematic policy where trial and error had been practised, by taking account of interdependence between instruments and by providing a quantitative indication of effects. Further, building on earlier work by Frisch distinguishing between various types of variables in relation to their role in policy models, Tinbergen demonstrated the connection between the analytical, or explanatory version and the policy, or normative version of economic models. In the analytical version, the policy targets were explained by other endogenous variables and by exogenous variables, which included the policy instruments. In the policy version the position of targets and instruments would be reversed (targets becoming exogenous and instruments endogenous variables) such that, in a well-behaved linear system, a solution requires only equality of the numbers of targets and instruments. This conclusion, which became known as the 'Tinbergen rule', brought an end to the popular misconception of a one-to-one correspondence between targets and instruments.

Development Economics

In reaction to his experiences during a trip to India in 1951, Tinbergen left the Central Planning Bureau in 1955 and moved to the field of development economics, more specifically the planning of the socio-economic development of low-income countries. Much earlier he had published a mathematical-statistical study of the theory of long-term economic growth (1942), but this had related only to industrialized countries. In the model technological progress had explicitly been included and the statistical tests (with data for

Great Britain, France, Germany and the United States from the decades before the First World War) already suggested that capital and labour growth could explain only a relatively small portion of the growth of production.

Characteristically, Tinbergen applied a quantitative, systematic policy approach to the development problem. This approach, which became known as 'planning-in-stages', distinguished macro, middle and micro stages, dealing with policy problems of private and public decision makers at the national, sectoral and project level, respectively (1967). In view of the difficult transportation conditions and the scarcity of skilled labour in developing countries, he subsequently added spatial and educational dimensions to the backbone of the planning-in-stages approach. He greatly simplified the calculation procedure for project evaluation by devising the semi-input-output method. This method was based on the notion that only the indirect effects emanating from sectors producing non-tradable (national) goods needed to be incorporated. At a time when computer capacity was still very limited, such a simplification was most useful. However, consistency between the micro stage and the other two levels was achieved only with the advent of computable general equilibrium models.

Tinbergen acted as adviser on matters related to economic development to the governments of Egypt, Turkey, Venezuela, Surinam, Indonesia and Pakistan, and he wrote studies for international organizations such as UNESCO and the OECD. As Chairman of the UN Committee on Development Planning from 1965 to 1972, he was involved with, among other things, the preparation of the UN Second Development Decade (1971–80).

Income Distribution

Tinbergen revisited the field of income distribution after his retirement as full professor (1972a; 1975). His approach, then as much as before, was inspired to a considerable extent by the positional-exchange criterion that had emerged from discussions in his student days with Paul Ehrenfest. According to this criterion a distribution of welfare could be considered fair when no one would wish to take another person's position. It was, for example, expressed in the individual welfare function Tinbergen proposed, which depended negatively on the difference (positive or negative) between the level of schooling required for a job and the actual schooling obtained by the person on this job. The notion that an income distribution is the outcome of a confrontation of demand and supply factors was another characteristic element of his approach. Thus, the development of a country's income distribution would be governed to a large extent by the process of technological innovation (a demand factor) and the rise of educational attainment levels (a supply factor). On the basis of material from the United States and The Netherlands from 1900 onwards, he found that this 'race' was mostly won by the rise in education, which resulted in more equitable distributions.

In his contributions to the field of income distribution – which concentrated on the remuneration of labour categories – he aimed to examine the effect of some unorthodox propositions. One such proposition was to consider the applicability of a capability tax which, as a lump sum tax, would be preferable to the familiar income tax. (Remarkably, this proposal ran counter to his finding that tax changes have a very slight impact on primary incomes, such that tax shifting would hardly be a problem.) Further, and true to his conviction that scientific progress and practical applications depend on quantitative tests of hypotheses, he treated welfare as measurable on the assumption that further progress in this area would be feasible. Assuming that workers move freely from one job to another so utility would be equalized, he derived an empirical relation expressing the connection between wage income on the one hand and attained schooling and the difference between attained and required schooling on the other. He then used this relation to compute an optimal or just distribution of income, tentatively relating to the situation in The Netherlands in the early 1960s. It would require very considerable shifts in income as compared with the actual situation.

International Economic Integration

Tinbergen's earliest work on international economic relations was still connected with national policymaking. Thus, his estimates of price elasticities of trade packages were meant to examine the effectiveness of a devaluation policy, where he emphasized the need to use long-term rather than short-term elasticities. His gravitation model (1962, Appendix VI) was a Newtonian approach to the explanation of bilateral trade flows which appeared to depend positively on the GNPs of the trade partners and negatively on the shipping distance separating them. It could be used to identify, among other things, the magnitude of potential trade lost to higher-than-average trade barriers, which impeded the efficient international division of labour he advocated in a number of studies written in the 1960s. Tinbergen (1954) applauded the international economic integration movement as it could remove trade barriers (which he dubbed negative economic integration) and could even result in new institutions for coordinated and centralized policymaking (positive economic integration). But he attached particular importance to the fact that economic integration would effectively reduce the probability of armed conflicts. From historical processes in Europe he derived a 'velocity of integration' which he hoped would remain positive until full integration at the regional and indeed the world level were achieved (1991a).

The Optimal Regime

His lifelong concern for (inter)national policymaking and, in that context, his special concern for the underdog resulted in a number of publications on the optimal economic order. In a deviation from his usual approach, Tinbergen emphasized in his Nobel Prize acceptance speech (1969) that the problem here consisted not of establishing the right mix of values of economic variables but of finding the proper set of institutions regarding the size and content of the public sector, the extent and content of (de)centralization of socio-economic decision making and therefore also of market regulation. He developed his ideas on the optimal order within a welfare-economic framework concerned with identifying the conditions that must be fulfilled to achieve maximum social welfare subject to the restrictions, such as production technologies, that apply in human society (1972b). In such a setting it would be useful to select the social welfare function at the beginning so as to limit the ethical possibilities in the subsequent analysis. The activities of the institutions would be described by a number of behaviour equations, the total of which should coincide with the conditions for optimal welfare. Tinbergen argued against rigidities, privileges, monopolies and insider-determined remunerations that bore no relation to marginal productivities, but he also rejected excessively generous social security systems that invited rent seeking.

In Tinbergen's view the interests of developing countries deserved separate attention in discussions on the optimal economic order. No country would accept within its borders an income inequality between groups of rich and poor citizens as could be found between rich and poor countries in the world. Not only must obstacles to exports from developing countries be removed; it would also be necessary to support these countries' development efforts by providing technical and financial aid. Tinbergen urged replacing the arbitrary UN target for international aid of 0.7 per cent of GNP of rich countries by the volume of aid that would be required for a harmonization of incomes within a predetermined number of years. He coordinated a study (1977) for the Club of Rome offering views on the international order, development aid, food production, the international division of labour, energy sources and raw materials, technological development, the environment and the arms race, among other things.

With the help of the theory of the optimal regime Tinbergen further sought to rid the confrontation between the Communist East and the Capitalist West of the dogmatic character that dominated world politics before the fall of Communism in 1989. Horrified by the prospect of nuclear warfare, he devoted a large part of his later years to a plea for a rational debate on the

pros and cons of both systems and for a stronger role of a reformed United Nations taking decisions that would incorporate international external effects (1990).

In Conclusion

Tinbergen's contribution to the economics discipline lies in his pioneering work in a number of different economic fields. He would not consider himself an expert even in these areas, would gladly admit that others who had come in after him had meanwhile gained a better understanding, and would move on to another area where another pressing social problem needed to be addressed. In his own words (1991b), 'solving the most urgent problems first' is what moved him most in his intellectual agenda.

He had little patience with studies lacking applicability to practical problems, and was not much impressed by scientific elegance for its own sake. His work discipline, punctuality and efficiency were exemplary. For an appointment, students and assistants he supervised would get seven minutes on the watch he would keep nearby. Still, Tinbergen also gave innumerable lectures for organizations and social action groups even of humble status.

His intense desire for a more humane world led him to put great trust in the benevolence and effectiveness of governments and international organizations, realizing that policies to overcome social problems would nearly always require the participation of public institutions. The latter's serious shortcomings in terms of management and governance were just another problem to be solved. He nursed a strong hope that people would behave more sensibly over time and learn to avoid the terrible conflicts that had caused so much suffering and devastation in the 20th century. It was for all these characteristics that Samuelson (2004, p. 153) described Tinbergen as 'a humanist saint'. Naturally, during his long life Tinbergen was often deeply disappointed. Still, his optimism never left him, if only because, as he said at an advanced age: 'I cannot afford to be pessimistic'.

See Also

- ▶ Development Economics
- ▶ Econometrics
- ▶ Foreign Aid
- ▶ Frisch, Ragnar Anton Kittel (1895–1973)
- ▶ Gravity Models
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Redistribution of Income and Wealth

Selected Works

1933. *Statistiek en Wiskunde in Dienst van het Konjunktuuronderzoek* [Statistics and mathematics in the service of business cycle research]. Inaugural Lecture, Netherlands School of Economics, Rotterdam, 4 October. Amsterdam: Arbeiderspers.

1936. An economic policy for 1936. In *Jan Tinbergen: Selected papers*, ed. L. Klaassen, L. Koyck and H. Witteveen. Amsterdam: North-Holland, 1959 (original in Dutch).

1939. *Statistical testing of business cycle theories. I: A method and its application to investment activity. II: Business cycles in the United States of America, 1919–1932*. Geneva: League of Nations.

1940. On a method of statistical business cycle research. A reply. *Economic Journal* 50: 141–54.

1942. Zur Theorie der Langfristigen Wirtschaftsentwicklung [On the theory of longterm economic growth]. *Weltwirtschaftliches Archiv* 55: 511–49.

1943. *Economische Bewegingsleer* [Economic dynamics]. Amsterdam: Noord-Hollandsche.

1950. (With J. Polak.) *The dynamics of business cycles: A study in economic fluctuations.* Chicago: University of Chicago Press.

1952. *On the theory of economic policy.* Amsterdam: North-Holland.

1954. *International economic integration.* Amsterdam: Elsevier.

1956. *Economic policy: Principles and design*. Amsterdam: North-Holland.

1962. *Shaping the world economy: Suggestions for an international economic policy.* New York: The Twentieth Century Fund.

1967. *Development planning*. London: World University Library.
1969. The use of models: experiences and prospects. In *Nobel lectures, economics 1969–1980*, ed. A. Lindbeck. Singapore: World Scientific Publishing Co., 1992. Online. Available at http://nobelprize.org/economics/laureates/1969/tinbergen-lecture.html. Accessed 14 June 2006.
1972a. An interdisciplinary approach to the measurement of utility or welfare. Fifth Geary lecture for The Economic and Social Research Institute, Dublin.
1972b. Some features of the optimum regime. In *Optimum social welfare and productivity*, ed. J. Tinbergen et al. New York: New York University Press.
1975. *Income distribution: Analysis and policies.* Amsterdam: North-Holland.
1977. *RIO: Reshaping the international order: A report to the Club of Rome*. New York: New American Library.
1979. Recollections of professional experiences. *Banca Nazionale del Lavoro Quarterly Review* 131, 331–60.
1990. *World security and equity.* Aldershot: Edward Elgar.
1991a. The velocity of integration. *De Economist* 139: 1–11.
1991b. Solving the most urgent problems first. In *Eminent economists: Their life philosophies*, ed. M. Szedberg. Cambridge and New York: Cambridge University Press.

Acknowledgments The authors gratefully acknowledge pertinent comments offered by the editors and Marcel Boumans.

Bibliography

Aftalion, A. 1927. The theory of economic cycles based on the capitalistic technique of production. *Review of Economic Statistics* 9: 165–170.
Duesenberry, J., G. Fromm, L. Klein, and E. Kuh. 1965. *The brookings quarterly econometric model of the United States*. Chicago: Rand McNally.
Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honour of Gustav Cassel*. London: Allen and Unwin.
Haavelmo, T. 1943. Statistical testing of business-cycle theories. *Review of Economic Statistics* 25: 13–18.
Hanau, A. 1928. Die Prognose der Schweinepreise [Forecasting the prices of pork], *Vierteljahreshefte zur Konjunkturforschung*, Sonderheft 18. Berlin: Institut für Konjunkturforschung, 1930.
Jevons, W. 1884. *Investigations in currency and finance*. London: Macmillan.
Jolink, A. 2003. *Jan Tinbergen: The statistical turn in Economics, 1903–1955*. Rotterdam: CHIMES.
Juglar, C. 1862. *Des Crises Commerciales et de Leur Retour Périodique en France, en Angleterre et aux Etats-Unies*. Paris: Guillaumin.
Keynes, J.M. 1919. *The economic consequences of the peace*. London: Macmillan.
Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.
Keynes, J.M. 1939. Professor Tinbergen's method. *Economic Journal* 49: 558–568.
Keynes, J.M. 1940. Comment. *Economic Journal* 50: 154–156.
Klein, L. 1950. *Economic fluctuations in the United States, 1921–1941*. New York: Wiley.
Klein, L., and A. Goldberger. 1955. *An econometric model of the United States, 1929–1952*. Amsterdam: North-Holland.
Kol, J., and P. de Wolff. 1993. Tinbergen's work: Change and continuity. *De Economist* 141: 1–28.
Leontief, W. 1950. *The structure of American economy.* New York: Oxford University Press.
Mitchell, W. 1913. *Business cycles and their causes*. Vol. III. Berkeley: University Memoirs.
Morgan, M. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.
Persons, W. 1919. An index of general business conditions. *Review of Economic Statistics* 1: 111–205.
Pronk, J. 2003. Tinbergen, idealist en inspirator. Lecture given on the occasion of the centennial of Tinbergen's birth. Erasmus University, Rotterdam, 9 April. Online. Available at http://www.janpronk.nl/index62.html. Accessed 14 June 2006.
Samuelson, P. 2004. Homage to Jan Tinbergen. In *Economics with a purpose: Tinbergen centennial issue*, ed. P. Cornelisse, H. van Dijk and H. Don. De Economist 152(2).
Slutzky, E. 1927. The summation of random causes as the source of cyclic processes. *The Problem of Economic Conditions* 3: 34–64 (English summary, 156–61).
Solow, R. 2004. Progress in economics since Tinbergen. *In Economics with a purpose: Tinbergen centennial issue*, ed. P. Cornelisse, H. van Dijk and H. Don. *De Economist* 152(2).
von Haberler, G. 1936. *Prosperity and depression: A theoretical analysis of cyclical fluctuations*. Geneva: League of Nations.
Yule, G. 1927. On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London, A* 226: 267–298.

Tintner, Gerhard (1907–1983)

Karl A. Fox

Keywords

Calculus of variations; Dynamic economic theory; Linear programming; Probability; Statistical inference; Stochastic programming; Technological risk; Tintner, G.; Uncertainty; Variate difference method

JEL Classifications

B31

Gerhard Tintner was born in Nuremberg, Germany, of Austrian parents, and educated in Vienna, completing his doctorate in economics, statistics and law at the University of Vienna in 1929. Tintner was much ahead of his time in important respects. First, he made early and significant contributions toward the development of a theory of behaviour under uncertainty (Tintner 1941a, b, 1942a, b, c). Second, he consistently stressed the need for a broad view of probability in the behavioural sciences and economics (Tintner 1960, 1968). His seminal article 'Foundations of Probability and Statistical Inference' (1949) started from Carnap's view of probability as *degree of confirmation* and raised issues some of which are now being debated in current reformulations of econometric methodology (Harper and Hooker 1976; Koch and Spizzichino 1982). Third, he firmly believed that the tools of modern disciplines such as cybernetics and system theory should be adapted and used to gain insight into individual and social behaviour, which is the basis of all applied economic models (Tintner and Sengupta 1972).

Tintner's first book (1935) was written as part of the programme of the Austrian Institute for Trade Cycle Research. In it Tintner applied Anderson's (1927) variate difference method to some 300 series of commodity prices from 1845 to 1914. Under certain assumptions, this method eliminated (most of) the random component from each series, leaving the systematic component for further study; Tintner (1940) presented a more complete statement of the method.

By 1935, Tintner had become enthusiastic about the work of the American mathematicians G.C. Evans (1922, 1924, 1930) and C.F. Roos (1925, 1934), who were applying calculus of variations to theoretical problems in economic dynamics. It appears that Tintner hoped to make a major breakthrough by extending the Evans–Roos approach. From 1936 to 1942 he published a series of brilliant articles (1936, 1937, 1938a, b, 1939, 1941a, b, 1942a, b, c) on such topics as maximization of utility over time, the theoretical derivation of dynamic demand curves, and the pure theory of production under technological risk and uncertainty; his last article of this type was 'A Note on Welfare Economics' (1946). Apparently these articles attracted little attention under the disturbed conditions of the time and were not consulted by the young economists who applied similar methods in the 1950s and 1960s to the theory of economic growth and uncertainty. Tintner's work on dynamic economic theory deserves a thorough reappraisal.

The early literature on linear programming dealt exclusively with the deterministic case. Tintner (1955) and Charnes and Cooper (1959) were the first to develop theories and methods for dealing with the various stochastic cases in which inputs, outputs, technical coefficients and/or constraints are subject to random disturbances. Tintner's development of an active approach to stochastic programming (as opposed to a passive approach) pointed the way to current research on self-tuning control combining both estimation and regulation (Sengupta 1985). Tintner's students and others also made important contributions to stochastic programming; by the late 1970s its literature included several hundred articles and a number of books – see, for example, Kolbin (1977), Tintner and Sengupta (1972), van Moeseke (1965), and Sengupta (1972, 1982).

A selected bibliography of Tintner's publications through 1967 is included in Fox, Sengupta and Narasimham (1969). Tintner spent the bulk of his career at Iowa State University (1937–62) and

the University of Southern California (1963–73). From 1973 until shortly before his death (in Vienna, 13 November 1983) he was professor of econometrics at the Technische Universität in Vienna and Honorary Professor at the University of Vienna. His textbooks (Tintner 1952, 1953) had considerable influence on the teaching of econometrics and he also published important articles on multivariate analysis, time series analysis and homogeneous systems in mathematical economics.

Selected Works

1935. *Prices in the trade cycle.* Vienna: Julius Springer.

1936. A note on distribution of income over time. *Econometrica* 4: 60–66.

1937. Monopoly over time. *Econometrica* 5: 160–170.

1938a. The maximization of utility over time. *Econometrica* 6: 154–158.

1938b. The theoretical derivation of dynamic demand curves. *Econometrica* 6: 375–380.

1939. Elasticities of expenditure in the dynamic theory of demand. *Econometrica* 7: 266–270.

1940. *The Variate difference method,* Cowles commission monograph no. 5. Bloomington: Principia Press.

1941a. The theory of choice under subjective risk and uncertainty. *Econometrica* 9: 298–304.

1941b. The pure theory of production under technological risk and uncertainty. *Econometrica* 9: 305–312.

1942a. A contribution to the non-static theory of choice. *Quarterly Journal of Economics* 56: 274–306.

1942b. A contribution to the non-static theory of production. In *Studies in mathematical economics and econometrics: In memory of Henry Schultz,* ed. O. Lange. et al. Chicago: University of Chicago Press.

1942c. The theory of production under nonstatic conditions. *Journal of Political Economy* 50: 645–667.

1946. A note on welfare economics. *Econometrica* 14: 69–78.

1949. Foundations of probability and statistical inference. *Journal of the Royal Statistical Society, Series A* (General) 112(Part III): 251–279.

1952. *Econometrics.* New York/London: Wiley/Chapman & Hall.

1953. *Mathematics and statistics for economists. New York: Rinehart.*

1955. Stochastic linear programming with applications to agricultural economics. In *Symposium on linear programming,* vol. 1. Washington, DC: National Bureau of Standards.

1960. *Handbuch der Ökonometrie.* Berlin: Springer.

1968. *Methodology of mathematical economics and econometrics.* Chicago: University of Chicago Press. Also issued as vol. 2, no. 6 of the *International encyclopedia of unified science.* Chicago: University of Chicago Press.

1972. (With J.K. Sengupta.) *Stochastic economics: Stochastic processes, control, and programming.* New York: Academic.

Bibliography

Anderson, O. 1927. On the logic of the decomposition of statistical series into separate components. *Journal of the Royal Statistical Society* 90(Part III): 548–569.

Charnes, A., and W.W. Cooper. 1959. Chance-constrained programming. *Management Science* 6 (1): 73–79.

Evans, G.C. 1922. A simple theory of competition. *American Mathematical Monthly* 29: 371–380.

Evans, G.C. 1924. The dynamics of monopoly. *American Mathematical Monthly* 31: 77–83.

Evans, G.C. 1930. *Mathematical introduction to economics*. New York: McGraw-Hill.

Fox, K.A., J.K. Sengupta, and G.V.L. Narasimham, eds. 1969. *Economic models, estimation and risk programming: Essays in honor of Gerhard Tintner*. Berlin/New York: Springer.

Harper, W.L., and C.A. Hooker, eds. 1976. *Foundations of probability theory, statistical inference and statistical theories of science*. Vol. 2. Dordrecht: D. Reidel.

Koch, G., and F. Spizzichino, eds. 1982. *Exchangeability in probability and statistics*. Amsterdam: North-Holland.

Kolbin, V.V. 1977. *Stochastic programming*. Dordrecht: D. Reidel.

van Moeseke, P. 1965. Stochastic linear programming. *Yale University Economic Essays* 5 (1): 197–253.

Roos, C.F. 1925. A mathematical theory of competition. *American Journal of Mathematics* 47: 163–175.

Roos, C.F. 1934. *Dynamic economics*, Cowles commission monograph. Vol. 1. Bloomington: Principia Press.
Sengupta, J.K. 1972. *Stochastic programming: Methods and applications*. Amsterdam: North-Holland.
Sengupta, J.K. 1982. *Decision models in stochastic programming*. Amsterdam: North-Holland.
Sengupta, J.K. 1985. *Information and efficiency in economic decision*. Dordrecht: Martinus Nijhoff.

Titmuss, Richard Morris (1907–1973)

Philippe Fontaine

Keywords
Arrow, K.; Social policy; Blood procurement; Buchanan, J.; Ethics and economics; Friedman, M.; Gifts; Hayek, F.; Institute of Economic Affairs; Invisible hand; Olson, M.; Socialism; Solow, R.; Titmuss, R

JEL Classifications
B31

A professor of social administration at the London School of Economics (LSE) from 1950 to his death, Richard Titmuss has often been depicted as an inept critic of economics. With *The Gift Relationship* (1970), however, he managed to attract the attention of leading economists. Robert Solow (1971) and Kenneth Arrow (1972), for instance, took pains to write lengthy review articles of what they and a number of their prominent peers, such as James Buchanan and Milton Friedman, regarded as a highly significant book. In a subject which has a solid tradition of confining ethical matters to its periphery and which has frequently resisted ideas emanating from other social sciences, it is paradoxical that a book of strong ethical inspiration and uncertain disciplinary origins attracted so much interest. One way out this paradox is perhaps to note that, starting with Mancur Olson's *The Logic of Collective Action* (1965), and accompanying the economic difficulties of the late 1960s, economists began to question the power of the invisible hand of the market in bringing about social cohesion. By the early 1970s the time was ripe for reconsidering the virtues of alternative coordinating mechanisms.

The context played a role in the reception of *The Gift Relationship*, but it was above all its subject that made the difference. Long before the book was published in early 1971, Titmuss (1959, 1963, 1968) had been a major figure in the debate over the welfare state, and in this capacity had criticized economists for their incomplete notion of what holds a society together and the unfortunate policy prescriptions they derived from it. With his new essay, however, he presented his own conception of social cohesion and social policy by contrasting the British system of blood procurement and distribution, based on free giving, to the partly commercialized US system. Metaphorically, the gift of blood illustrated the consolidation of the social bond, while its sale stood for social collapse. In other words, Titmuss pointed to the dangers of society's increasing commercialization.

The son of a small farmer and his wife of less humble origins, Titmuss lacked formal schooling beyond the age of 14, and when he joined the Eugenics Society in 1937 it was hardly expected he would end up in academe. Yet, after participating in several of the society's research projects, including those of the Population Investigation Committee, then headed by Alexander Carr-Saunders, Director of the LSE, he made a name for himself in academic circles. Subsequently, the historian Keith Hancock approached him to work as historian of the Cabinet office, as a result of which Titmuss produced the monumental *Problems of Social Policy* (1950). Not only did the book secure him the chair in Social Administration at the LSE but it also illustrated the importance of the experience of the Second World War in shaping his 'vision of good conduct as generalised obligation' (see Reisman 2004). Based on values of solidarity and social duty, this vision may well have been suitable to describe the wartime and immediate post-war British society. From the mid-1960s, however, it became clear that it was incomplete, as was the vision of those economists who regarded

the market as the main, if not exclusive, coordinating mechanism in society.

It is these economists' ideas and the applications suggested by their allies at the Institute of Economic Affairs (IEA), the London-based think tank inspired by Hayek, that Titmuss fought from the late 1960s to 1973 (see Fontaine 2002). He was of the opinion that 'social growth' was more important than its economic counterpart, and that a 'socialist' social policy – not the invisible hand of the market – was essential to social cohesion. These two ideas were intimately connected. Titmuss used the former to show that the economists' social indicators were inadequate: the economy could grow economically and still regress socially because negative externalities supplanted positive ones. Likewise, 'socialist' social policies stimulated ethical behaviour, which generated positive externalities and averted negative externalities, whereas 'private' social policies, as envisaged by the IEA, favoured commercialism, which neglected positive externalities and underestimated negative externalities.

While Titmuss's criticism that economists relied too much on the invisible hand of the market may have actually applied only to a few of them (Solow 1971), his thesis that too much commercialism undermines the social bond concerned them all. Most economists remained unconvinced, but, faced with Titmuss's emphasis on the role of the gift in so unusual a setting as impersonal interactions, where they typically saw selfishness as reigning supreme, the economists were forced to contemplate the possibility that 'a world of giving may actually increase efficiency in the operation of the economic system' (Arrow 1972, p. 351).

See Also

► Altruism, History of the Concept

Selected Works

1950. *Problems of social policy*. London: HMSO.
1959. *Essays on the welfare State*. New Haven: Yale University Press.
1963. Ethics and economics of medical care. *Medical Care* 1: 16–22.
1968. *Commitment to welfare*. London: Allen and Unwin.
1970. *The gift relationship: From human blood to social policy*. London: Allen and Unwin.

Bibliography

Arrow, K. 1972. Gifts and exchanges. *Philosophy and Public Affairs* 1: 343–362.
Fontaine, P. 2002. Blood, politics and social science: Richard Titmuss and the Institute of Economic Affairs, 1957–1973. *Isis* 93: 401–434.
Olson, M. *1965*. *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
Reisman, D. 2004. Richard Titmuss: Welfare as good conduct. *European Journal of Political Economy* 20: 771–794.
Solow, R. 1971. Blood and thunder. *Yale Law Journal* 80: 1696–1711.

Tobin, James (1918–2002)

Donald D. Hester

Abstract

James Tobin was a brilliant economist and the leading proponent of Keynesian economics in the second half of the 20th century. He greatly advanced understanding of financial institutions and monetary theory and policy. He stressed the importance of asset holdings and wealth on consumer spending. He also argued that 'q', the ratio of a firm's market value to the replacement value of its assets, was an important determinant of its investment decisions. He made major contributions to econometric methods, international economics, the theory of growth and business cycles, and policies designed to improve the welfare of minorities and the poor.

Keywords

Absolute income hypothesis; Allais, M.; American Economic Association; Animal spirits; Banking industry; Baumol, W.; Bonds; Economic growth; Financial intermediation; Fiscal policy; Fleming, J.; Floating exchange rate; Friedman, M.; Government debt; Household portfolios; Income elasticity of demand; Income velocity of money; Interest rate elasticity; Investment decisions; IS–LM model; Keynes, J. M.; Keynesianism; Leisure; Liquidity preference; Lucas, R.; Maximum likelihood; Measure of economic welfare (MEW); Monetarism; Monetary policy; Monetary theory; Monetary transmission mechanism; Multicollinearity; Mundell, R.; National income accounts; Negative income tax; Non-accelerating inflationary rate of unemployment (NAIRU); Permanent-income hypothesis; Phillips curve; Pollution; Portfolio balance; Portfolio demand for money; Rational expectations hypothesis; Rationing; Relative income hypothesis; Ricardian equivalence theorem; Solow, R.; Stocks and flows; Swan, T.; Technology; Traditional vs modern; Tobin tax; Tobin, J.; Tobin's q; Tobit model; Transactions demand for money; Wage rigidity

JEL Classifications

B31

James Tobin was a major contributor to economic science and macroeconomic analysis who received the Nobel Memorial Prize in Economic Science in 1981 for 'his analysis of financial markets and their relations to expenditure decisions, employment, production, and prices'. He received a BA in 1939, an MA in 1940, and a Ph.D. in 1947 from Harvard University. (Between 1941 and 1945 Tobin was an officer on a destroyer in the US Navy.) He made many early penetrating theoretical and empirical explorations of the underpinnings of Keynes's *General Theory of Employment, Interest, and Money* (1936). Throughout his long career he was a vigorous participant in the design of fiscal and monetary policies and in actively promoting their implementation. He was on the faculty of Yale University from 1950, and from 1957 until his death he held a prestigious Sterling professorship there.

He published a large number of influential papers in professional journals and other places, and his impact on economics can only be hinted at in this unavoidably incomplete article. Buiter (2003) has published a much longer but still incomplete survey of Tobin's work, which consists of about 500 articles over a period of 60 years. Many of his papers are accessible in a series of volumes of his collected works (1971, 1975, 1982, 1996a) in addition to their original places of publication.

Apart from his dissertation, his earliest empirical evaluations of the Keynesian approach (1947, 1951), were influential informal studies that respectively examined the interest elasticity of the demand for money and the underpinnings of the consumption function. The latter showed that evidence for Duesenberry's (1949) 'relative income hypothesis' was not persuasive when compared with a wealth- augmented variation of Keynes's absolute income hypothesis, when individuals were broken down into savers and dissavers and variations in the cost of living were taken into account. The importance of wealth and portfolio composition on behaviour would be a major theme in Tobin's work in subsequent decades.

Perhaps his most important applied econometrics article (1950) examined the demand for food in the United States. It refined techniques that had been used in his dissertation, which analysed the consumption function. By combining estimates of income elasticities obtained from cross-sectional budget data with price elasticities estimated from time series, Tobin attempted to avoid the problem of multicollinearity that often arises when one uses time series data exclusively. He focused on food rather than aggregate consumption in this paper in order to avoid problems associated with purchases of durable goods. This paper has stood the test of time remarkably well and was celebrated in a special issue of the *Journal of Applied Econometrics* (1997).

Shortly after finishing this paper, Tobin (1952a) and Tobin and H. S. Houthakker, (1951; 1952) completed an elegant series of papers on the theory of rationing and its implications for econometric analysis.

Household Balance Sheets and Spending

Upon arriving at Yale, Tobin developed a distinctive empirical approach to household spending and portfolio behaviour that was foreshadowed by an illuminating survey article, 'Asset Holdings and Spending Decisions' (1952b). He critically reviewed attempts to include money, liquid assets, and government debt holdings by households as predictors of consumption in both macroeconomic and microeconomic studies. He argued that such asset holdings are best viewed as part of a dynamic optimization process; people accumulate liquid assets in preparation for executing spending programmes. Liquid assets cannot sensibly be interpreted as exogenous determinants of consumption.

An early empirical study (1957) analysed the relation between consumer holdings of debt, liquid assets and durable goods and decisions to acquire more of each of them. He reported that high levels of debt discouraged acquisition of more debt and durable goods, and in an illuminating diagram suggested that individuals with different attributes tend to have different desired mixtures of debt and liquid assets, which can be interpreted as target levels of 'portfolio balance'. A second empirical study, coauthored with H. W. Watts (1960) extended this portfolio balance approach to broad collections of assets and liabilities.

Tobin made a very important econometric methods contribution that was a consequence of his work with household assets. He noted that many household assets took on zero values until a certain threshold was crossed. Linear regressions attempting to represent relations, say, between a household's income and the value of its car would be severely misspecified if some observations in a sample had no car.

In (1958a) he developed an ingenious probit-regression technique (subsequently named 'tobit' by A. S. Goldberger), which allowed maximum likelihood techniques to be used to obtain consistent estimates of such relationships. The technique continues to be widely used.

A Dynamic Aggregative Model and the Effects of Money on Economic Growth

Portfolio balance assumed a central role in a very innovative model (1955) where Tobin managed to combine endogenous economic fluctuations and economic growth. This paper, which Tobin viewed as his favourite (see Breit and Spencer 1986, p. 128), is Keynesian in that money wages may be inflexible and may lead to unemployed labour. An important innovation in the model was a linear homogeneous production function that allowed for substitution between capital and labour, as in the nearly contemporaneous economic growth models of Solow (1956) and Swan (1956). However, its major new contribution was the formal introduction of portfolio balance and money. Changes in money are equal to the government deficit in this model. Wealth consists of physical capital and the real money supply. Momentary price equilibrium exists when the price level allows wealth holders to be satisfied with the mix of capital and real money holdings in their portfolios. Because investors are assumed to be risk averse, portfolio equilibrium does not require that the negative expected rate of change of price equals the expected marginal productivity of capital. Depending upon price expectations, monetary expansion may affect the rate of capital formation positively or negatively. The paper briefly explores how technical progress relates to price changes and growth.

Introducing inflexible money wage rates leads to a variety of results that include inflation, deflation, secular economic growth or stagnation, and economic cycles, which can be seen when two summary relations describing labour market balance and portfolio balance are examined. Recovery from cyclical slumps may occur if extreme

conditions cause money wage rates to become flexible or if capital depreciates sufficiently.

In this model and in its specializations to economic growth (1965b, 1968), money and government debt are indistinguishable. In an important unpublished manuscript dating from 1958, which eventually was published in 1998, Tobin addressed this distinction and presented an elegant analysis of how monetary policy worked through banks and how changes in the composition of government debt affected capital and growth. In Tobin (1961) and in his contribution to the Commission on Money and Credit (1963a), Tobin drew on this manuscript to analyse debt and monetary policies and their relation to the return on capital. The reasons for the long delay in publication of the manuscript are unclear, but surely part of the explanation was that Tobin was a member of the Kennedy administration's Council of Economic Advisors for 18 months beginning in January 1961.

The Demand for Money

Tobin published two papers that were narrowly focused on the demand for money (1956, 1958b). The first was an inventory theoretic model of the transactions demand for money that had close antecedents in Allais (1947) and Baumol (1952). These three papers cast doubt on the constancy of the income velocity of money, Y/ M, by arguing that an individual's income elasticity of demand for money is likely to be about 0.5 and the interest rate elasticity about – 0.5. As the theory predicted, the post-war income velocity of money in the United States rose as per capita income and interest rates rose.

The second paper was a path-breaking effort that proposed a novel explanation for the portfolio demand for money. It was written at about the time Markowitz was writing his monograph (1959) on portfolio selection at Yale, but had a very different orientation. For simplicity Tobin assumed cash had a riskless rate of return. He introduced uncertainty about the rate of return on a second asset, and explored how an investor should split his portfolio between cash and the risky asset over a single planning period. His goal was to eliminate an unsatisfactory assumption about interest rate expectations of investors that underlay Keynes's exposition of liquidity preference. He assumed either that investors' utility functions were quadratic or that the distribution of the risky rate of return was normal, and derived the optimal mixture of the risky asset and cash. Although his discussion was not error-free, as was pointed out by Borch (1969) and Feldstein (1969), it provided a foundation from which a large literature in finance developed. Initially Tobin assumed the second asset was a consol, in the spirit of Keynes's discussion. Then he showed that the second asset could be a linear combination of a set of risky assets and that investors' portfolio problem could often be reduced to a mixture of that combination and cash. He was the first to recognize 'two fund' separation, which would play a major role in the theory of finance.

Financial Intermediation and Policy

Tobin's unpublished manuscript included a discussion of commercial banks and their role in the transmission of monetary policy, topics that he extensively developed in the 1960s. An early unpublished paper (1959) examined how financial intermediaries in general responded to monetary controls; it would be extensively revised and appear with the same title in a rigorous discussion of monetary policy, coauthored with William Brainard (1963). Tobin and Brainard argued that a monetary policy action's effect can be measured with the required rate of return on real capital; restrictive (easing) monetary policy will raise (lower) this rate. They assumed that all assets and liabilities are gross substitutes. Thus, for assets an increase in the rate of return on one asset increases the demand for it and does not increase the demand for any other asset, and analogously for liabilities. They analysed how a monetary policy-induced change in currency affects the required rate of return in regimes with controlled and uncontrolled intermediaries, where controls consist of reserve requirements and interest rate ceilings on deposit liabilities. The controls

may strengthen the efficacy of monetary policy, but it works when they are absent. Tobin (1963b) provided an accessible intuitive discussion of the responses of banks and other intermediaries to monetary policy actions.

Tobin (1969) presented a general equilibrium interpretation of the overall approach and emphasized the importance of 'q', the ratio of the market value of a firm to the replacement value of its assets, a central focus of his approach to monetary policy and a prominent variable in subsequent empirical studies of the investment decision. Intuitively, when q is greater (less) than unity, stockholders gain (lose) when a firm undertakes net new investment. *Ceteris paribus*, a high real rate of interest (and thus restrictive monetary policy) is likely to reduce market values and thus discourage new investment. Tobin and Brainard (1977) reported a very ambitious attempt to measure average q using panel data and introduced the important distinction between marginal and average q, a topic that was subsequently thoroughly developed by Hayashi (1982). While acknowledging that q was an endogenous variable partly determined by animal spirits and investor expectations, they argued that an estimate of q for firms should be included among the variables guiding monetary policy.

Tobin and Brainard (1968) reported simulation experiments using a hypothetical model that illustrated the central role of q and the importance of taking into account cross equation restrictions. Drawing on well-known arguments from the theory of consumer demand, they suggested that empirical models of financial markets often were misspecified because investigators ignored adding-up constraints on parameters across equations on interest rate, income and lagged variables. They argued that such requirements were especially important in dynamic specifications, where often the only lag in an asset demand equation was the asset's value in a previous period; in principle, all lagged asset variables in a well-specified portfolio adjustment model must appear in each asset demand equation. Implicitly, these two Tobin and Brainard articles indicate the enormous sensitivity of models to arbitrary accounting conventions and the unavoidable errors in variables that such conventions lead to.

Monetary Theory and Policy

Beginning with his review (1965a) of Friedman and Schwartz's *Monetary History of the United States*, Tobin increasingly assumed the role of defending the Keynesian approach to macroeconomics against an emerging monetarist formulation that was led by Milton Friedman. Tobin had been a major contributor to the highly successful policies of the Kennedy and Johnson administration policies of the early 1960s, but was very critical of the latter's financing of the Vietnamese war (1981, pp. 357, 360). As the decade wore on, the Keynesian approach came increasingly under attack. For example, Milton Friedman's presidential address (1968) to the American Economic Association, which had been preceded by Phelps (1967) with the same idea, questioned the existence of a stationary Phillips curve. Friedman argued that workers would increasingly seek redress from inflation if unemployment were below the non-accelerating inflationary rate of unemployment (NAIRU, where the price-inflation Phillips curve crosses the abscissa). If a non-stationary Phillips curve led to accelerating rates of inflation, prices would no longer be sticky as the Keynesian approach assumed. Friedman's alternative approach argued that money demand was a function of permanent income and that empirical support for the monetarist model came from the fact that in time series money led income, which suggested that money was a causal element. Tobin and Swan (1969) reported empirical relations between permanent income and money that did not support the monetarist model. Tobin (1970) reported simple theoretical models that illustrated that one could not infer from lead–lag relationships in time series of money and income whether a model was fundamentally Keynesian or monetarist. Space limitations prevent a full and fair exposition of the debate, which continued with varying levels of intensity for many years; see, for example, Tobin (1993b, 1995).

The controversies led to ambitious efforts to see whether fiscal and monetary policy effects in the static IS–LM model were likely to carry over in a dynamic long- run framework. Tobin and Buiter (1976) and Tobin (1979) examined this question

formally using elegant techniques and were not able to conclude in general that the effects of policy were either transitory or long-lasting in models with a stationary state. They were able to describe some model specifications with predictable effects in the long run and discussed their intuitive plausibility. Tobin and Buiter (1980) examined the same issue in the context of a growth model, with similarly inconclusive results. All the same, these papers are valuable examples of how to analyse such questions.

Tobin provided a relatively complete statement of his views on the theory of macroeconomic policy in his three Yrjö Jahnsson lectures and a related Paish Lecture given in England that appear in Tobin (1980). His first Jahnsson lecture considered how price level changes affected economic activity, focusing on the Pigou effect, the Keynes effect, and Fisher's debt-deflation hypothesis. He believed that the Fisher hypothesis won over the Pigou effect in the short and medium run. He denied that Keynes had demonstrated an underemployment equilibrium, but accepted the IS–LM model as a reasonable guide to policy in the short run.

His second Jahnsson lecture attacked both Friedman's views about a shifting Phillips curve that are discussed above and Lucas's (1972) rational expectations hypothesis. He argued that the latter was more radical than Friedman's discussion because it denied even the short-run efficacy of monetary policy. Tobin explained that, while everyone agrees that expectations are important, expectations are highly diverse in an economy and based on different information sets; therefore, they are not likely to be unbiased. Further, he denied the Lucas contention that labour and product markets are always being cleared at existing wages and prices, and criticized as ad hoc Lucas's specification 'about the information available to buyers and sellers' (Tobin 1980, p. 42).

His third Jahnsson lecture addressed an obvious deficiency of the short-run IS–LM model, namely, that the capital stock and other measures of wealth are assumed to be constant even though investment is occurring and debt is being issued. He sketched out a model that integrates flow-of-funds accounts with variables in the IS–LM model, which derived from his papers with Buiter. An especially ambitious attempt at constructing such a model is reported in Backus, Brainard, Smith and Tobin (1980).

The Paish lecture examined the plausibility of Ricardian equivalence – the idea that there is no difference resulting from decisions to finance an increase in government spending by printing money and/or raising taxes, and/or issuing government debt. Robert Barro (1974) had resurrected and developed this thesis, which denies that government deficits have an effect because households take into account the fact that increased issues of debt would require an increase in future taxes, which would be needed to service and retire the debt. While the Barro argument might be true in a simple hypothetical world where household dynasties are infinitely lived and each household in a generation behaves so as to fully take into account effects on its counterparts in other generations, Tobin argued that demographic events like no heirs, generational myopia, and financial conditions such as liquidity constraints, possibly resulting from imperfect capital markets, and wealth and other distorting taxes invalidated it.

Economic Development and the Measurement of Economic Activity

As a visiting professor at the University of Nairobi in the 1972–3 academic year, Tobin wrote two instructive papers that use the linear activity analysis model. The first (1974a) examined a model with two technologies, traditional and modern (more productive), where unemployment was possible in an economy with a constant saving rate and where capital in each of the technologies had possibly different depreciation rates. Depending upon parameter values, he showed that it could be optimal to invest capital in the traditional rather than the modern technology, in both the short and the long runs. In (1974b) he studied the effects of expulsion of alien residents and expropriation of their assets. Depending upon the compensation, if any, for expropriating the assets of the former residents, the linear model

yielded predictions about how remaining citizens would fare. Tobin did not analyse the ethical justification, if any, for expulsion and expropriation, but nevertheless provided a very insightful discussion about who might gain and lose in situations where expropriation occurred.

William Nordhaus and Tobin in a very ambitious paper entitled 'Is Growth Obsolete?' (1972) proposed a primitive measure of economic welfare (MEW) for the United States. They argued that the national income accounts measure production and are not appropriate for analysing welfare, for several reasons. The accounts do not measure the flow of services from the stock of durable goods or the value of human capital acquired through education and training, nor do they account for the depletion of minerals, environmental degradation, and the disamenities of urbanization. The national income accounts include all consumer and government expenditures. Nordhaus and Tobin viewed many of these as 'instrumental' expenditures that are regrettably necessary for welfare. Examples include commuting costs, police services, national defence and sanitation. They subtracted such expenditures from their measure, which is expressed on a per capita basis, and made an imputation for the flow of leisure. The authors argued that technical progress and capital formation have more than compensated for the depletion of natural resources so far and are guardedly optimistic that this will continue. However, they stated that the effects of pollutants on melting polar ice caps warrant a higher priority in research and that pollution is a problem because it is not priced to reflect social costs. While the possibility of catastrophic global disturbances cannot be excluded, their conclusion was that growth is not obsolete.

International Economics

Tobin and Braga de Macedo (1980) modified the standard paradigm of Mundell (1961) and Fleming (1962) of how monetary policy works in a floating exchange rate system by introducing the exchange rate in asset demand functions and assuming that all assets are gross substitutes. There are three assets – cash, foreign holdings and bonds – because bonds and capital are conventionally assumed to be perfect substitutes. In the case of a small open economy they showed that with these modifications fiscal policy affects the economy's level of real income, although the effect cannot be signed without information about some parameters in this variation of a standard IS–LM model. This contrasts with the Mundell and Fleming models where fiscal policy is impotent in a pure floating exchange rate system.

The authors then expanded the small open-economy model to have four assets, by assuming that bonds and capital are not perfect substitutes. They continued to assume that all assets are gross substitutes and employed a discrete time specification to analyse some of the consequences of asset accumulation. This second model describes a single period, but beginning-of-period stocks and within-period flows yield complex effects that unfortunately cannot be summarized adequately in the present article. However, it is usually the case that fiscal policy again has non-zero effects in a flexible exchange regime. They also examined fiscal policy in a two- country, flexible-exchange world, where neither is small and in each country bonds and capital are perfect substitutes. The rich analytical framework of this paper is applied insightfully in Tobin (1993a) and partly underlies a simulation exercise by Tobin and Brainard (1992).

Tobin (1978) argued that countries should impose a tax on purchases of financial instruments denominated in another country's currency in order to allow some autonomy in setting domestic stabilization policies. This controversial 'Tobin tax' proposal to throw sand into the gears of international finance first appeared in Tobin (1974c, pp. 88–92), achieved wide approval outside the United States, and was analysed in ul Haq, Kaul and Grunberg (1996).

Overview and Summary

As stated at the outset, one cannot do justice to Tobin's extraordinary career in a short article. However, before closing, it is important to

acknowledge his extensive contributions to public debate on policy that appear in Tobin (1966; 1996b), and especially on welfare and inequality that are reprinted in part in Tobin (1982, pp. 497–624). He was a forceful advocate of the negative income tax, reducing inequality, and improving the economic condition of minorities in the United States, especially during the turbulent 1960s and 1970s.

See Also

- IS–LM
- Keynesian Revolution
- Monetary and Fiscal Policy Overview
- Monetary Transmission Mechanism
- Natural Rate of Unemployment
- Rationing
- Ricardian Equivalence Theorem
- Tobin's q
- Tobit Model

Selected Works

1947. Liquidity preference and monetary policy. *Review of Economics and Statistics* 29(2): 124–131.

A statistical demand function for food in the U.S. A. *Journal of the Royal Statistical Society* 113, Series A, Part II: 113–141.

Relative income, absolute income, and saving. In *Money, trade, and economic growth (essays in Honor of John Henry Williams)*. New York: Macmillan.

(With H. Houthakker.) The effects of rationing on demand elasticities. *Review of Economic Studies* 18(3): 1–14.

1952a. A survey of the theory of rationing. *Econometrica* 20: 521–553.

1952b. Asset holdings and spending decisions. *American Economic Review* 42(2): 109–123.

(With H. Houthakker.) Estimates of the free demand for rationed foodstuffs. *Economic Journal* 62: 103–118.

A dynamic aggregative model. *Journal of Political Economy* 63(2): 103–115.

The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38(3): 241–247.

Consumer debt and spending: Some evidence from analysis of a survey. *Consumer installment credit*. Part II, vol. 1. Washington, DC: US Government Printing Office.

1958a. Estimation of relationships for limited dependent variables. *Econometrica* 26 (1): 24–36.

1958b. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.

Financial intermediaries and the effectiveness of monetary controls. Discussion Paper No. 63 (January), Cowles Foundation, Yale University.

(With H. Watts.) Consumer expenditures and the capital account. In *Proceedings of the conference on consumption and saving,* ed. I. Friend and R. Jones. Philadelphia: University of Pennsylvania Press.

Money, capital, and other stores of value. *American Economic Review* 51(2): 26–37.

1963a. An essay on the principles of debt management. In *Fiscal and debt management policies.* Englewood Cliffs: Prentice Hall.

1963b. Commercial banks as creators of 'money'. In *Banking and monetary studies,* ed. D. Carson. Homewood: Richard D. Irwin.

1963. (With W. Brainard.) Financial intermediaries and the effectiveness of monetary controls. *American Economic Review* 53: 383–400.

1965a. The monetary interpretation of history. *American Economic Review* 55: 464–485.

1965b. Money and economic growth. *Econometrica* 35: 671–684.

1966. *National Economic Policy.* New Haven: Yale University Press.

Notes on optimal monetary growth. *Journal of Political Economy* 76: 833–859.

(With W. Brainard.) Pitfalls in financial model building. *American Economic Review* 58(2): 99–122.

A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1(1): 15–29.

(With C. Swan.) Money and permanent income: Some empirical tests. *American Economic Review* 59: 285–295.

Money and income: Post hoc ergo propter hoc? *Quarterly Journal of Economics* 84: 301–317.

Essays in economics: Macroeconomics. Chicago: Markham Publishing.

(With W. Nordhaus.) Is growth obsolete? In *Economic growth: Fiftieth anniversary colloquium V*, National Bureau of Economic Research. New York: Columbia University Press.

1974a. Technological development and employment. *Eastern Africa Economic Review* 6(1): 1–26.

1974b. Notes on the economic theory of expulsion and expropriation. *Journal of Development Economics* 1(1): 7–18.

1974c. *The new economics, one decade older.* Princeton: Princeton University Press.

Essays in economics: Consumption and econometrics. Amsterdam: North-Holland.

(With W. Buiter.) Long-run effects of fiscal and monetary policy on aggregate demand. In *Monetarism: Studies in monetary economics*, vol. 1, ed. J. Stein. Amsterdam: North-Holland.

(With W. Brainard.) Asset markets and the cost of capital. In *Economic progress, private values and public policy*, ed. R. Nelson and B. Balassa. Amsterdam: North-Holland.

A proposal for international monetary reform. *Eastern Economic Review* 4 (3–4): 153–159.

Deficit spending and crowding out in shorter and longer runs. In *Theory for economic efficiency: Essays in Honor of Abba P. Lerner,* ed. H. Greenfield et al. Cambridge: MIT Press.

Asset accumulation and economic activity: *Reflections on contemporary macroeconomic theory.* Yrjö Jahnsson Lectures. Chicago: University of Chicago Press.

1980. (With D. Backus, W. Brainard, and G. Smith.) A model of US financial and non-financial behavior. *Journal of Money, Credit, and Banking* 12: 259–293.

(With W. Buiter.) Fiscal and monetary policies, capital formation, and economic activity. In *The government and capital formation*, ed. G. von Furstenberg. Cambridge: Ballinger.

(With J. Braga de Macedo.) The short-run macroeconomics of floating exchange rates: An exposition. In *Flexible exchange rates and the balance of payments: Essays in memory of Egon Sohmen,* ed. J. Chipman and C. Kindleberger. Amsterdam: North-Holland.

Reflections inspired by proposed constitutional restrictions on fiscal policy. In *Economic regulation: Essays in honor of James R. Nelson*, ed. K. Boyer and W. Shepherd. East Lansing: Michigan State University.

Essays in economics: Theory and policy. Cambridge: MIT Press.

1992. (With W. Brainard.) On the internationalization of portfolios. *Oxford Economic Papers* 44: 553–565.

1993a. Policies and exchange rates: A simple analytical framework. In *Japan, Europe, and international financial markets: Analytical and empirical perspectives*, ed. R. Sato, R. Levich and R. Ramachandran. Cambridge: Cambridge University Press.

1993b. Price flexibility and output stability: An old Keynesian view. *Journal of Economic Perspectives* 7(1): 45–65.

1995. The natural rate as new classical macroeconomics. In *The natural rate of unemployment: Reflections on 25 years of the hypothesis,* ed. R. Cross. Cambridge: Cambridge University Press.

1996a. *Essays in economics: National and international*. Cambridge: MIT Press.

1996b. *Full employment and growth: Further Keynesian essays on policy.* Cheltenham: Edward Elgar.

1998. (With S. Golub.) *Money, credit, and capital.* Boston: Irwin/McGraw-Hill.

Bibliography

Allais, M. 1947. *Economie et intérêt*. Paris: Impremerie Nationale.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Baumol, W. 1952. The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics* 66: 545–556.

Borch, K. 1969. A note on uncertainty and indifference curves. *Review of Economic Studies* 36: 1–4.

Breit, W., and Spencer, R., eds. 1986. Tobin. In *Lives of the laureates.* Cambridge: MIT Press.
Buiter, W. 2003. James Tobin: An appreciation of his contribution to economics. *Economic Journal* 113: 585–631.
Duesenberry, J. 1949. *Income, saving, and the theory of consumer behavior.* Cambridge: Harvard University Press.
Feldstein, M. 1969. Mean-variance analysis in the theory of liquidity preference and portfolio selection. *Review of Economic Studies* 36: 5–12.
Fleming, J. 1962. Domestic financial policies under fixed and under floating exchange rates. *International Monetary Fund Staff Papers* 9: 369–379.
Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1): 1–17.
Hayashi, F. 1982. Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica* 50: 213–224.
Lucas, R. 1972. Econometric testing of the natural rate hypothesis. In *The economics of price determination conference*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.
Keynes, J.M. 1936. *General theory of employment, interest, and money.* New York: Macmillan.
Magnus, J., and M. Morgan, eds. 1997. The experiment in applied econometrics. *Journal of Applied Econometrics* 12(5): 459–662.
Markowitz, H. 1959. *Portfolio Selection: Efficient diversification of investments. Cowles Foundation Monograph 16.* New York: John Wiley and Sons.
Mundell, R. 1961. Flexible exchange rates and employment policy. *Canadian Journal of Economics and Political Science* 27: 509–517.
Phelps, E. 1967. Phillips curves, expectations of inflation, and optimal unemployment over time. *Economica* 34: 254–281.
Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Swan, T. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
ul Haq, M., I. Kaul, and I. Grunberg, eds. 1996. *The Tobin Tax: Coping with financial volatility.* New York: Oxford University Press.

Tobin's q

Gary Smith

Abstract

Tobin's q is the ratio of the market value of a firm to the replacement cost of its assets, a statistic that depends on the firm's profitability and financial markets' required rate of return. Although there are a variety of measurement issues, including the distinction between marginal and average q, Tobin's q can be used to predict investment spending or to control for a firm's current and future profitability in empirical studies of corporate structure and behaviour.

Keywords

Accounting conventions; Adjustment costs; Corporate finance; Depreciation; Heterogeneous capital; Inflation; Investment decisions; Keynes, J. M; Monopoly rents; Rate of return; Takeovers; Technological progress; Tobin, J; Tobin's q

JEL Classifications

E22

A standard tenet of corporate finance is that the retention of earnings to finance expansion raises a stock's price if the rate of return on these investments, ρ, is larger than shareholders' required return on their stock, R. For example, an investment that costs \$1 million is worth more than \$1 million to stockholders if $\rho > R$. This principle suggests that whether a firm's stock sells for a premium or a discount relative to the cost of its assets depends on ρ versus R and, further, that a firm's investment decisions ought to depend on a comparison of ρ with R.

The shareholders' required return is determined in financial markets by shareholders pricing stock to yield an anticipated return that is competitive with comparable investments they might make. This determination of required returns is one of the primary ways in which financial markets affect real economic activity. When interest rates fall, required stock returns decline, making it more likely that $\rho > R$, so that the profits from prospective investments are sufficient to make these investments attractive for firms that care about their shareholders.

But where do firms (or economic forecasters) see these shareholder required returns? It is

straightforward to calculate the yield to maturity on a bond by determining the discount rate that equates the present value of the promised cash flow to the market price. There are no comparable calculations for stocks because the cash flow to investors is unknown. An ingenious alternative, proposed by James Tobin (Brainard and Tobin 1968; Tobin 1969), is to look at stock prices.

Specifically, Tobin argues that we should look at how financial markets value a firm relative to the replacement cost of the firm's assets:

$$q = \frac{\text{market value}}{\text{replacement cost}}$$

The numerator and denominator of Tobin's q can be aggregate market value and replacement cost or, equivalently, price per share and assets per share. If a firm has debts, these can be included in the numerator.

What Determines q?

How can assets be valued in financial markets at other than their replacement cost? Assets are of value to shareholders only to the extent that they generate profits. It matters not at all that a factory cost $1 billion to build if it doesn't make a cent of profits. For a factory to be worth what it costs shareholders, it must earn the shareholder's required rate of return.

For a simple example, suppose that a hamburger chain has no debt and pays out all earnings as dividends. Assume also that a new restaurant costs $1 million to build and is expected to earn a constant 20 per cent profit ($\rho = 0.20$), $200,000 a year for ever. The value that financial markets place on the $200,000 annual cash flow depends on how highly hamburger earnings are valued. If Treasury bonds yield five per cent, perhaps stock in risky hamburger restaurants is priced to yield ten per cent ($R = 0.10$). Because the anticipated dividends are a constant $200,000, the market value of the restaurant is

$$V = \frac{\$200,000}{0.10} = \$2,000,000$$

Valued at $2 million, the $200,000 annual dividend provides stockholders their requisite ten per cent return.

In this case, the market value of the restaurant is twice its construction cost: $q = 2$. Stockholders welcome this investment, because the use of $1 million in potential dividends to build the restaurant provides $2 million in market value. The underlying reason is that the restaurant's 20 per cent profit rate is larger than the market's ten per cent required rate of return.

If, on the other hand, shareholders' required rate of return is 25 per cent, then

$$V = \frac{\$200,00}{0.25} = \$800,000$$

Now $q = 0{:}8$ and the construction of the restaurant will be to the detriment of shareholders. Because the restaurant earns only 20 per cent on its cost, the stock must be valued at less than the asset's cost in order to provide shareholders their 25 per cent required return.

Implications for Investment Decisions

A connection between business investment spending and market value relative to cost was pointed out many years ago by John Maynard Keynes (1936, p. 151):

> [T]he daily revaluations of the Stock Exchange, though they are primarily made to facilitate transfers of old investments between one individual and another, inevitably exert a decisive influence on the rate of current investment. For there is no sense in building up a new enterprise at a cost greater than that at which a similar existing enterprise can be purchased; whilst there is an inducement to spend on a new project what may seem an extravagant sum, if it can be floated on the stock exchange at an immediate profit.

Similarly, Tobin argues that a firm should invest in new buildings and equipment if the stock market will value the project at more than its cost (that is, if the project's q is greater than 1). If the market value is larger than the cost, shareholders prefer that the firm make this investment rather than distribute its cost as dividends, gladly giving up a dollar of dividends in exchange for a two-dollar increase in the value of their stock.

Put more plainly, the appropriate question a firm should ask is whether, if it were to sell shares in its new venture, it could raise enough money to cover the project's cost. It can if the value of Tobin's q is larger than 1, but not otherwise. Thus Tobin's q provides a barometer of the incentives for business investment.

Similarly, the firm should compare the price it can get for selling its existing assets with the value that financial markets place on these assets. If the market value is less than the sale price ($q < 1$), the firm is worth more dead than alive, and it should sell off its assets and distribute the proceeds either through dividends or share repurchases. A low market value relative to replacement cost may also motivate takeover bids, since an outside group may profit by purchasing enough stock to gain control of a company and then liquidating its assets.

Marginal and Average q

Why don't firms immediately exploit any differences between market value and replacement costs, thereby causing q to return to 1 instantaneously? Three, sometimes related, explanations involve convex adjustment costs, monopoly rents and heterogeneous capital (Lucas and Prescott 1971; Mussa 1977; Smith 1981; Hayashi 1982; Abel 1983; Erickson and Whited 2000; Gomes 2001). A consideration of these issues requires a distinction between average q, the aggregate market value and replacement cost of a firm's assets, and marginal q, the change in a firm's market value resulting from a specific investment relative to the cost of that investment.

An observed average q that is greater than 1 might accurately reflect a company's substantial profits, but further investment may be restrained by convex adjustment costs that cause the marginal q for a large expansion to be less than 1. It might be prohibitively expensive for a hamburger chain to triple its size in a year. A related strand of literature (summarized by Hubbard 1998) investigates how investment by financially constrained firms may be less sensitive to q and more sensitive to the firm's cash flow.

Similarly, a firm might earn monopoly rents that cause average q to exceed 1, but have a marginal q that is less than 1 because new investments would erode these rents. A hamburger restaurant with a patented secret formula might not want to open a competing restaurant next door that would lure away customers. The grower of an exotic fruit may not want to flood the market with produce that could only be sold by lowering prices.

Finally, prospective ventures might be quite different from the firm's existing operations, with different rates of return and with risks that command different required returns. A tobacco company acquires a snack food company; a yogurt maker enters the bottled water market; a software company enters the video game market. In each case, marginal q might be quite different from average q.

A further complication is that observed market values presumably take into account not only existing assets but also future investments anticipated by the market. Suppose, for example, that a firm currently has assets with a replacement cost and market value that both equal $100 million and is planning to make an investment that will cost $20 million and have a market value of $30 million. Average q for the firm's current assets is 1 and marginal q for this investment is 1.5. If the stock market takes into account this projected investment, the current market value is increased by $10 million (discounted somewhat to the extent the value added will occur in the future and if there is uncertainty about whether the investment will be made). Thus observed average q reflects both the profitability of the current capital stock and the perceived profitability of the firm's future opportunities, overstating the former and understating the latter.

Estimates of q

The book value reported by firms is a rough proxy, a starting point, for estimating the replacement cost of a firm's assets. Accounting conventions value assets at historical cost, with no adjustments for subsequent cost increases, and depreciate assets according to accounting conventions rather than true economic depreciation. Inflation can cause book values to understate the replacement cost of assets (think of real estate); technological progress can cause the reverse (think of computers). The market value of a firm's equity is

commonly estimated by multiplying the market price of a firm's stock by the number of shares outstanding; data on the market value of a firm's preferred stock and debt are not so easily obtained because databases generally record the book values reported by firms in their balance sheets.

A variety of often complex procedures have been employed to estimate the market value of a firm's debts and the replacement cost of its assets (Brainard et al. 1980; Lindenberg and Ross 1981; Lewellen and Badrinath 1997; Lee and Tompkins 1999), while other authors argue that more readily available book values provide sufficiently accurate approximations (Chung and Pruitt 1994; Perfect and Wiles 1994).

Because of these measurement errors, when q is used as an explanatory variable in a regression model, least squares estimates of the coefficient of q will be biased towards zero and estimates of the coefficients of other explanatory variables may be biased towards zero or away from zero. For example, in a model that uses a firm's q and current cash flow to predict investment spending, the coefficient of q will be biased downward (towards zero) and the coefficient of cash flow may be biased upward. Erickson and Whited (2000) propose sophisticated estimators to deal with measurement error and obtain relatively large estimates of the relationship between q and investment.

Another issue is whether we should use the market's valuation or the firm's internal valuation of prospective investments, since the firm may have better information about the projected cash flows and speculative stock market noise can causes market prices to wander from fundamental values (Morck et al. 1990; Blanchard et al. 1993). Several authors have proposed creative ways of estimating marginal q from information available to managers (Abel and Blanchard 1986; Gilchrist and Himmelberg 1995) or from stock analysts' earnings forecasts (Cummins et al. 2006; Bond and Cummins 2000). Gentry and Mayer (2006) apply the q model to real estate investment trusts (REITs) and find that the use of appraised value in place of accounting-based replacement cost increases the estimated empirical relationship between REIT investment and q.

Uses of q

Empirical studies using Tobin's q initially focused on either explaining q (Lindenberg and Ross 1981; Salinger 1984) or using q to predict investment spending (von Furstenberg 1977; Summers 1981; Hayashi 1982), but have since broadened to include many issues in corporate finance that hold investment opportunities constant. For example, if we want to use cross-section data to see whether dividend policy affects the value of a firm, we need to control for each firm's profitability. Thus q has been used in studies of the effects of managerial equity ownership (Morck et al. 1988; McConnell and Servaes 1990), the size of a company's board of directors (Yermack 1996), corporate diversification (Berger and Ofek 1995; Rajan et al. 2000); and dividend changes (Lang and Litzenberger 1989; Denis et al. 1994). For similar reasons, Tobin's q has been used to hold investment opportunities constant while investigating the determinants of capital structure (Titman and Wessels 1988), leveraged buyouts (Opler and Titman 1993), and takeovers (Lang et al. 1989; Servaes 1991).

Tobin's q will no doubt be used in many other empirical studies of corporate structure and behaviour because it circumvents the unresolved issue of how to estimate shareholders' risk-adjusted required return by looking directly at observable market prices, which incorporate both the cash flow expectations of investors and the required returns they use to discount this anticipated cash flow.

See Also

- ▶ Monetary Transmission Mechanism
- ▶ Profit and Profit Theory
- ▶ Tobin, James (1918–2002)

Bibliography

Abel, A. 1983. Optimal investment under uncertainty. *American Economic Review* 73: 228–233.

Abel, A., and O. Blanchard. 1986. The present value of profits and cyclical movements in investment. *Econometrica* 54: 249–273.

Berger, P., and E. Ofek. 1995. Diversification's effect on firm value. *Journal of Financial Economics* 37: 39–65.

Blanchard, O., C. Rhee, and L. Summers. 1993. The stock market, profit, and investment. *Quarterly Journal of Economics* 108: 115–136.

Bond, S., and J. Cummins. 2000. The stock market and investment in the new economy: Some tangible facts and intangible fictions. *Brookings Papers on Economic Activity* 2000(1): 61–124.

Brainard, W., J. Shoven, and L. Weiss. 1980. The financial valuation of the return to capital. *Brookings Papers on Economic Activity* 1980(2): 453–502.

Brainard, W., and J. Tobin. 1968. Pitfalls in financial model-building. *American Economic Review* 58: 99–122.

Chung, K., and S. Pruitt. 1994. A simple approximation of Tobin's q. *Financial Management* 23: 70–74.

Cummins, J., K. Hassett, and S. Oliner. 2006. Investment behavior, observable expectations, internal funds. *American Economic Review* 96: 796–810.

Denis, D., D. Denis, and A. Sarin. 1994. The information content of dividend changes: Cash flow signaling, overinvestment, and dividend clienteles. *Journal of Financial and Quantitative Analysis* 29: 567–587.

Erickson, T., and T. Whited. 2000. Measurement error and the relationship between investment and q. *Journal of Political Economy* 108: 1027–1057.

Gentry, W. and Mayer, C. 2006. What can we learn about the sensitivity of investment to stock prices with a better measure of Tobin's q? Working paper, Columbia University.

Gilchrist, S., and C. Himmelberg. 1995. Evidence on the role of cash flow in reduced-form investment equations. *Journal of Monetary Economics* 36: 541–572.

Gomes, J. 2001. Financing investment. *American Economic Review* 91: 1263–1285.

Hayashi, F. 1982. Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica* 50: 213–224.

Hubbard, R. 1998. Capital market imperfections and investment. *Journal of Economic Literature* 36: 193–225.

Keynes, J.M. 1936. *The general theory of employment, interest, and money.* London: Macmillan.

Lang, L., and R. Litzenberger. 1989. Dividend announcements: Cash flow signaling vs. free cash flow hypothesis? *Journal of Financial Economics* 24: 181–191.

Lang, L., R. Stulz, and R. Walkling. 1989. Managerial performance, Tobin's q, and the gains from successful tender offers. *Journal of Financial Economics* 24: 137–154.

Lee, D., and J. Tompkins. 1999. A modified version of the Lewellen and Badrinath measure of Tobin's q. *Financial Management* 28: 20–31.

Lewellen, W., and S. Badrinath. 1997. On the measurement of Tobin's q. *Journal of Financial Economics* 44: 77–122.

Lindenberg, E., and S. Ross. 1981. Tobin's q ratio and industrial organization. *Journal of Business* 54: 1–32.

Lucas Jr., R., and E. Prescott. 1971. Investment under uncertainty. *Econometrica* 39: 659–681.

McConnell, J., and H. Servaes. 1990. Additional evidence on equity ownership and corporate value. *Journal of Financial Economics* 27: 595–612.

Morck, R., A. Shleifer, and R. Vishny. 1988. Management ownership and market valuation: An empirical analysis. *Journal of Financial Economics* 20: 293–316.

Morck, R., A. Shleifer, and R. Vishny. 1990. The stock market and investment: Is the market a sideshow? *Brookings Papers on Economic Activity* 1990(2): 157–202.

Mussa, M. 1977. External and internal adjustment costs and the theory of aggregate and firm investment. *Economica* 44: 163–178.

Opler, T., and S. Titman. 1993. The determinants of leveraged buyout activity: Free cash flow vs. financial distress costs. *Journal of Finance* 48: 1985–1999.

Perfect, S., and K. Wiles. 1994. Alternative constructions of Tobin's q: An empirical comparison. *Journal of Empirical Finance* 1: 313–341.

Rajan, R., H. Servaes, and L. Zingales. 2000. The diversification discount and inefficient investment. *Journal of Finance* 55: 35–80.

Salinger, M. 1984. Tobin's q, unionization, and the concentration-profits relationship. *RAND Journal of Economics* 15: 159–170.

Servaes, H. 1991. Tobin's q and the gains from takeovers. *Journal of Finance* 46: 409–419.

Smith, G. 1981. Investment and q in a stock valuation model. *Southern Economic Journal* 47: 1007–1020.

Summers, L. 1981. Taxation and corporate investment: A q-theory approach. *Brookings Papers on Economic Activity* 1981(1): 67–127.

Titman, S., and R. Wessels. 1988. The determinants of capital structure choice. *Journal of Finance* 43: 1–19.

Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1: 15–29.

von Furstenberg, G. 1977. Corporate investment: Does market valuation matter in the aggregate? *Brookings Papers on Economic Activity* 1977(2): 347–397.

Yermack, D. 1996. Higher market valuation of companies with a small board of directors. *Journal of Financial Economics* 40: 185–212.

Tobit Model

Jean-Marc Robin

Abstract

Tobit models are used to model variables subject to exogenous censoring. For example, duration data cannot be observed longer than

the survey period; hours of work cannot be observed negative although an individual might be better off consuming more leisure time than is available. This article reviews a list of econometric techniques to estimate Tobit models. Maximum likelihood, Heckman's two-stage estimator, and Powell's trimmed least squares are successively addressed.

Keywords

Censored regression model; Endogenous regressors; Heteroskedasticity; Labour supply; Maximum likelihood; Non-normal errors; Ordinary least squares; Probit model; Tobit model; Trimmed least squares; Two-stage estimation

JEL Classifications

C25; C24

The Tobit model, or censored regression model, is useful to learn about the conditional distribution of a variable y^* given a vector of regressors x, when y^* is observed only if it is above or below some known threshold (censoring). In the original model of Tobin (1958), for example, the dependent variable was expenditures on durables, and values below zero are not observed.

Censoring models state that the observed dependent variable y follows from the latent variable y^* as

$$y = \max\{y^*, 0\},$$

where we have assumed a censoring of the form $y^* > 0$ without loss of generality because, for any given top or bottom threshold a, it is always possible to change y^* into $\pm(y^* - a)$.

Censoring may either be a property of the sample or a property of the population. For example, top-coding of earnings in the Current Population Survey (CPS) generates censoring in a way that is independent of individual decisions. In contradistinction, the zero purchases of Tobin's households *are* individual decisions. This type of censoring is usually modelled as a corner solution of a decision-theoretic model. For example, the labour supply model predicts that the number of hours worked by a person is equal to the interior solution of the consumption-and-leisure utility maximization problem, if it is greater than zero; it is zero otherwise. (See Pudney 1989, for a survey of the economics and econometrics of corner solutions.)

The relationship between the latent variable y^* and regressors x is assumed linear:

$$y^* = x^T\beta + u,$$

where β is a vector of parameters, $x^{\mathrm{T}}\beta$ denotes the scalar product of x and β ($^{\mathrm{T}}$ is the transpose operator), and u is a residual component with cumulative distribution function (cdf) F conditional on x. We assume that the distribution of u given x, that is F, is continuous. It hence has a density $f = F'$.

The Tobit model corresponds to the particular case of $F(u) = \Phi\left(\frac{u}{\sigma}\right)$ where Φ denotes the cdf of the standard normal distribution $N(0,1)$, and σ^2 is the variance of u (that is $u \sim N(0,\sigma^2)$).

The distribution of *y* given *x*

Let $G(y|x)$ denote the cdf of the observation y given x. The distribution of y is not continuous. It has a mass point at 0. The probability mass at 0 is

$$\begin{aligned} G(0|x) - \Pr\{y = 0|x\} &= \Pr\{y^* \le 0|x\} \\ &= \Pr\{u \le -x^T\beta|x\} \\ &= F(-x^T\beta). \end{aligned}$$

Notice that $F(-x^{\mathrm{T}}\beta) = 1 - F(x^{\mathrm{T}}\beta)$ if u has a symmetric distribution.

Any positive observation $y > 0$ is necessarily such that $y = x^{\mathrm{T}}\beta + u$. Therefore, the cdf of the observed outcome at $y > 0$ given x is equal to the cdf of u at $y - x^{\mathrm{T}}\beta$:

$$G(y|x) = F(y - x^T\beta), \forall y > 0.$$

The density of any observation $y > 0$ given x is

$$g(y|x) = \frac{\partial G(y|x)}{\partial y} = f(y - x^T\beta).$$

Notice that, since 0 is a mass point of the distribution of y, its density at 0 can be defined as

$$g(0|x) = G(0|x) - G(0^-|x) = G(0|x),$$

where $G(0^-|x) = \lim_{y\to 0-} G(y|x) = 0$. (The probability density function, pdf, of a distribution or random variable is defined relative to a particular measure. Continuous variables admit a density with respect to the Lebesgue measure. Discrete distributions admit a density with respect to the counting measure. One can also define a density function for mixed discrete-continuous distributions with respect to mixtures of the Lebesgue measure and the counting measure.)

In the case of the Tobit model, $f(u) = \frac{1}{\sigma}\phi\left(\frac{u}{\sigma}\right)$, where $\phi(v) = \Phi'(v) = \frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}$ is the density of the standard normal distribution. So,

$$g(0|x) = G(0|x) = F(-x^T\beta) = \Phi\left(\frac{x_i^T\beta}{\sigma}\right),$$
$$g(y|x) = f(y - x^T\beta) = \frac{1}{\sigma}\phi\left(\frac{y - x^T\beta}{\sigma}\right), \forall y > 0.$$

Moments of *y* Given *x*

Two conditional moments of y are of particular interest: E($y|x$)and E($y|x, y > 0$). First, notice that

$$y = \max\{y^*, 0\} \geq y^*$$

implies that

$$\mathbb{E}[y|x] \geq \mathbb{E}[y^*|x]$$

and

$$\mathbb{E}[y|x, y > 0] = \frac{\mathbb{E}[y|x]}{\Pr\{y > 0|x\}} \geq \mathbb{E}[y|x].$$

So both E($y|x$) and E($y|x, y > 0$) overestimate the first moment of the variable of interest, that is E($y^*|x$).

Specifically,

$$\begin{aligned}\mathbb{E}[y|x] &= \mathbb{E}[\max\{y^*|0\}|x] \\ &= \mathbb{E}[\max\{x^T\beta + u, 0\}|x] \\ &= \int_{x^T\beta}^{+\infty} (x^T\beta + u)f(u)\mathrm{d}u \\ &= x^T\beta(1 - F(-x^T\beta)) + \int_{x^T\beta}^{+\infty} uf(u)\mathrm{d}u\end{aligned}$$

and

$$\mathbb{E}[y|x, y > 0] = \frac{\mathbb{E}[y|x]}{\Pr\{y > 0|x\}} = x^T\beta + \lambda\left(x^T\beta\right)$$

with

$$\lambda(z) = \frac{\int_z^{+\infty} uf(u)\mathrm{d}u}{1 - F(z)}$$

In the particular case of the Tobit model, $\phi'(v) = -v\ \phi\ (v)$. It thus follows that $\lambda(v) = \sigma\frac{\phi\left(\frac{v}{\sigma}\right)}{\Phi\left(\frac{v}{\sigma}\right)}$. Notice that $\frac{\phi}{\Phi}$ is the inverse Mills ratio of the standard normal distribution.

Ordinary Least Squares

Let $\{(y_i, x_i), i = 1, \ldots, N\}$ be an i.i.d. random sample of observations. Regressing y_i on x_i for the uncensored observations i such that $y_i > 0$ does not yield a consistent estimator of β because of the omitted variable $\lambda(x^{\mathrm{T}}\ \beta)$ which is correlated with the regressors x_i.

For the Tobit model, a two-stage estimation procedure can be devised.

1. First, Estimate a Probit Model for $d_i \equiv 1\{y_i > 0\}$ (= 1 if $y_i > 0$ and = 0 Otherwise):

$$\Pr\{d_i = 1|x_i\} = \Phi\left(x_i^T c\right),$$

with $c = \frac{\beta}{\sigma}$ Let $\hat{c}$ be the Probit estimator of c.

2. Regress y_i on x_i and the inverse Mills ratio $\frac{\phi\left(x_i^T\hat{c}\right)}{\Phi\left(x_i^T\hat{c}\right)}$ by OLS. This yields a consistent estimator of β and σ.

Two remarks are in order. First, as any multistage estimation procedure, the OLS estimator of the second stage has $\frac{\phi\left(x_i^T\hat{c}\right)}{\Phi\left(x_i^T\hat{c}\right)}$ instead of $\frac{\phi\left(x_i^Tc\right)}{\Phi\left(x_i^Tc\right)}$.

The measurement error $\frac{\phi(x_i^T\hat{c})}{\Phi(x_i^T\hat{c})} - \frac{\phi(x_i^T c)}{\Phi(x_i^T c)}$ tends to 0 when N tends to infinity. So the OLS estimator of (β,σ) is asymptotically unbiased and consistent. However, its asymptotic variance has to be corrected for the statistical error on parameter c.

Second, the first stage requires knowing the entire distribution of u_i. It is therefore not clear why one would want to use this two-stage procedure instead of maximum likelihood (ML), which is efficient.

Maximum Likelihood

The likelihood of one observation y_i conditional on x_i is $g(y_i|x_i)$. The conditional sample log-likelihood is then

$$\begin{aligned} L_N &= \sum_{i=1}^{N} \ln g(y_i|x_i) \\ &= \sum_{i=1}^{N} \left\{ d_i \ln f\left(y_i - x_i^T\beta\right) + (1+d_i)\ln F\left(-x_i^T\beta\right) \right\}. \end{aligned}$$

Under standard regularity conditions, the values of β and any other parameters of F that maximize the log-likelihood L_N are root-N consistent and asymptotically normal and efficient.

For the Tobit model, we obtain

$$L_N = \sum_{i=1}^{N} (1-d_i)\ln\left(1 - \Phi\left(\frac{x_i^T\beta}{\sigma}\right)\right) - N_+ \ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N} d_i\left(y_i - x_i^T\beta\right)^2$$

where $N_+ = \sum_{i=1}^{N}(1+d_i)$ is the number of uncensored observations.

It is useful to change (β,σ) into $(c,s) = \left(\frac{\beta}{\sigma}, \frac{1}{\sigma}\right)$ because

$$L_N = \sum_{i=1}^{N} (1-d_i)\ln\left(1 - \Phi\left(x_i^T c\right)\right) + N_+ \ln s - \frac{1}{2}\sum_{i=1}^{N} d_i\left(s y_i - x_i^T c\right)^2$$

is strictly concave with respect to (c,s). Maximizing L_N with respect to (c,s) is easy and fast using standard gradient algorithms. One can then use the delta method to recover an estimate of the asymptotic variance of $\left(\hat{\beta},\hat{\sigma}\right) = \left(\frac{\hat{c}}{\hat{s}}, \frac{1}{\hat{s}}\right)$.

Consistency of ML obviously rests on the model being well specified. Non-normal errors and heteroskedasticity (when homoskedastic, normal errors are assumed) lead to inconsistent estimates.

Trimmed Least Squares

Powell's (1986) symmetrically trimmed least squares is a simple consistent estimator that is consistent under the assumption that the distribution of u_i is symmetric. It can yet be non-normal or heteroskedastic.

The idea is to replace y_i by $2x_i'\beta$ when $y_i \geq 2x_i'\beta$, if $x_i'\beta > 0$ and drop all observations such that $x_i'\beta \leq 0$ from the sample, as no symmetric trimming is possible in this case. In effect, let

$$\tilde{y}_i = \min\left\{y_i, 2x_i^T\beta\right\} = x_i^T\beta + \tilde{u}_i$$

where

$$\tilde{u}_i = \begin{cases} -x_i^T\beta & \text{if} \quad u_i \leq -x_i^T\beta \\ u_i & \text{if} \quad -x_i^T\beta < u_i \leq x_i^T\beta \\ x_i^T\beta & \text{if} \quad x_i^T\beta < u_i \end{cases}$$

As u_i has a symmetric distribution conditional on x_i, then so does $\tilde{u}_i$. Hence,

$$\mathbb{E}(\tilde{y}_i|\,\tilde{x}_i) = x_i^T\beta.$$

The trimmed least squares estimator is obtained by iterating the following sequential procedure until convergence. Start with an initial value β_0 for β. For example, regress y_i on x_i on the uncensored sample. If, at iteration p, one has obtained a value β_p for β, then compute β_{p+1} by regressing $\tilde{y}_i(\beta_p) = \min\left\{y_i, 2x_i^T\beta_p\right\}$ on x_i using the sample of observations i such that $x_i'\beta_p > 0$.

Endogenous Regressors

In the standard labour supply model, the observed dependent variable y_i is the actual number of hours worked by individual i, and the latent variable y_i^* is the interior solution to a utility maximization problem. This interior solution depends on the individual's wage, w_i, and other variables x_i such as non-labour income or education and age:

$$y_i^* = x_i^T \beta + \alpha w_i + u_i.$$

The residual u_i captures unobserved heterogeneity factors influencing the trade-off between consumption and leisure. It is usually understood that wages w_i and unobserved taste shifters u_i are correlated across individuals: $\mathrm{Cov}(w_i, u_i) \neq 0$.

Suppose that w_i and x_i are both observed when $y_i = 0$. The following simple control-function procedure can apply to solve the endogeneity problem. Suppose that there exists a vector z_i of instruments such that

$$w_i = z_i^T \gamma + v_i,$$

with $\mathrm{Cov}(z_i, v_i) = 0$. Suppose also that

$$u_i = \rho v_i + \varepsilon_i,$$

with ε_i normal $N(0,\sigma^2)$ conditional on x_i, z_i and v_i. This will be the case in particular if u_i and v_i are jointly normal conditional on x_i and z_i.

Then, the following two-stage procedure produces consistent estimators of β, α and ρ.

1. Regress w_i on z_i by OLS and compute residuals $\hat{v}_i$.
2. Estimate the Tobit model

$$y_i = \max\left\{x_i^T \beta + \alpha w_i + \rho \hat{v}_i + \eta_i, 0\right\},$$

assuming $\eta_i = \varepsilon_i - \rho(\hat{v}_i - v_i)$ normally distributed, by ML or other appropriate method.

One can test for the exogeneity of w_i by testing for $\rho = 0$ with a standard t-test. If the null hypothesis is rejected, then this two-stage procedure yields consistent estimates, but correct asymptotic standard errors, accounting for the approximation of v_i by $\hat{v}_i$, require a specific calculation.

Finally, if w_i is not observed when $y_i = 0$, which is the case for wages of not employed individuals, this procedure does not work. Heckman (1974) assumed joint normality of (u_i, v_i) and applied maximum likelihood to (y_i, w_i), $i = 1,...,N$, conditional on exogenous variables.

See Also

- Endogeneity and Exogeneity
- Logit Models of Individual Choice
- Maximum Likelihood
- Selection Bias and Self-Selection
- Two-Stage Least Squares and the k-Class Estimator

Bibliography

Blundell, R.W., and R.J. Smith. 1986. An exogeneity test for a simultaneous equation Tobit model with an application to labor supply. *Econometrica* 54: 679–686.

Blundell, R.W., and R.J. Smith. 1989. Estimation in a class of simultaneous equation limited dependent variable models. *Review of Economic Studies* 56: 37–57.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42: 679–694.

Powell, J.L. 1986. Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* 54: 1435–1460.

Pudney, S. 1989. *Modelling individual choice: The econometrics of corners, kinks and holes*. Oxford: Blackwell.

Tobin, J. 1958. Estimation for relationships with limited dependent variables. *Econometrica* 26(2): 24–36.

T

Tocqueville, Alexis Charles Henri Clérel de (1805–1859)

J. Goodwin

Keywords

Civil society; Democracy; Equality of conditions; Individualism; Liberty; Physiocracy; Tocqueville, A. C. H. C. de

JEL Classifications
B31

Alexis de Tocqueville was born at Verneuil, in Normandy, France, on 29 July 1805. In 1831 he journeyed to the United States with his friend Gustave de Beaumont to study the American penal system. He then wrote *Democracy in America,* the first volume of which appeared in 1835, the second in 1840. Tocqueville was a member of the French Chamber of Deputies and served briefly as Minister of Foreign Affairs in the republic established after the Revolution of 1848. The events of this period are recounted in his *Recollections* (1893). Tocqueville was among those arrested during the *coup d'état* of Louis Napoleon on 2 December 1851, and he subsequently retired from public life. Tocqueville devoted his last years to a major study of the French Revolution, although he completed only the first volume before his death. This appeared as *The Old Régime and the French Revolution* in 1856. He died at Cannes on 16 April 1859.

Tocqueville was interested in the political, cultural, and, to a lesser extent, economic consequences of 'democracy', by which he meant not representative government or political arrangements of any sort, but 'equality of conditions'. (John Stuart Mill would argue that Tocqueville had confounded the effects of 'democracy' with the tendencies of modern commercial society.) By equality of conditions, Tocqueville meant neither the absence of classes nor mere equality of opportunity, but something like rough social equality, including, especially, the absence of the legally prescribed hierarchy of social groups characteristic of 'aristocratic' societies.

Tocqueville's major intellectual, not to say political, preoccupation was discovering how 'liberty' might be preserved under democratic conditions. By 'liberty' Tocqueville meant above all the local control and administration of a community's common affairs by a politically engaged and civic-minded populace. He was thus a strong critic of both the administrative centralization of the state and the narrow, self-interested 'individualism' of bourgeois society. This explains Tocqueville's appeal to those on both the right and left of the political spectrum.

Tocqueville's search for the institutional and ideological supports of liberty under democratic conditions was the ulterior purpose of his journey to the United States, a country which had managed to combine democracy and liberty. In *Democracy in America,* Tocqueville analysed a number of factors which he believed helped to maintain political liberty in the United States, including administrative decentralization, the profusion of voluntary associations, and what he termed 'self-interest properly understood', that is, a disposition to devote part of one's time and wealth to the good of the community. *The Old Régime and the French Revolution,* by contrast, explored the failure of France's revolutionary transition to democracy to produce a stable liberal regime; this Tocqueville attributed principally to the immense administrative centralization of the pre-revolutionary period and the consequent degradation of French political culture.

Tocqueville's reflections on economic matters are few. In fact, he once 'confessed' to Nassau Senior that he 'was insufficiently informed on this important portion of human science'. In *The Old Régime,* however, Tocqueville did not hesitate to criticize the Physiocrats, whom he believed perhaps best represented the abstract and utopian type of intellectual nourished by the illiberal environment of pre-revolutionary France. Tocqueville thought that the Physiocrats lacked any concern for political, as opposed to economic, liberty, offering only the 'intellectual panacea' of universal education.

They were for abolishing all hierarchies, all class distinctions, all differences or rank, and the nation was to be composed of individuals almost exactly alike and unconditionally equal. In this undiscriminated mass was to reside, theoretically, the sovereign power; yet it was to be carefully deprived of any means of controlling or even supervising the activities of its own government.

Tocqueville's reflections on economic matters in *Democracy in America* comprise only a

few pages of that voluminous work. Tocqueville argued that rents tend to rise and the terms of leases to shorten in democracies owing to the dissolution of the close, customary relationship between landlord and tenant and its replacement by the impersonal contract. He also thought that democratic conditions made it easier for workmen to combine and pressure their employers for higher wages; he thus argued that 'a slow, progressive rise in wages is one of the general laws characteristic of democratic societies'. At the same time, Tocqueville feared that the very richest industrialists could wait out strikes and force permanently lower wages on their workers. In fact, Tocqueville believed that a dangerous business or industrial 'aristocracy' might arise within the womb of democratic society. However, this potential aristocracy was not to be greatly feared, Tocqueville thought, since industrialists seldom look beyond their own interests and share no common traditions or corporate spirit; still, Tocqueville warned, 'if ever again permanent inequality of conditions and aristocracy make their way into the world, it will have been by that door that they entered'.

Selected Works

1835. *Democracy in America,* vol. 1, ed. J.P. Mayer. Trans. G. Lawrence. New York: Doubleday, 1969.
1840. *Democracy in America,* vol. 2, ed. J.P. Mayer. Trans. G. Lawrence. New York: Doubleday, 1969.
1856. *The old régime and the French revolution.* Trans. S. Gilbert from the 4th French ed. of 1858. New York: Doubleday, 1955.
1872. *Correspondence and conversations of Alexis de Tocqueville with Nassau William Senior, 1834–1859*. 2nd ed., ed. M.C.M. Simpson. Reprinted, New York: A.M. Kelley, 1968.
1893. *Recollections,* ed. J.P. Mayer and A.P. Kerr. Trans. G. Lawrence. New York: Doubleday, 1970.

Tooke, Thomas (1774–1858)

Massimo Pivetti

Keywords

Bank Charter Act (1844); Bank of England; Banking School; Convertibility; Cost of production; Currency School; Endogenous money; Free trade movement; Political Economy Club; Price level; Rate of interest; Tooke, T.; Torrens, R.; Wicksell, J. G. K.

JEL Classifications

B31

Thomas Tooke, the leading member of the Banking School, was born at St Petersburg in 1774, the eldest son of William Tooke, historian of Russia and man of letters, at that time chaplain to the English church at St Petersburg. Not a professional scientist but an active man of business of comfortable social standing, Thomas was successively a partner in the London firms of Stephen Thornton & Co. and Astell, Tooke & Thornton, Russian merchants, and was governor of the Royal Exchange Corporation and chairman of the St Katharine's Dock Company. In 1802 he married Priscilla Combe, by whom he had three sons.

As an early supporter of the free trade movement, he drew up the Merchants' Petition of the City of London, which contained the statement of the principles of free trade and was presented to the House of Commons in May 1820. He gave evidence on monetary questions before several parliamentary committees, from the Resumption Committees of 1819 to the Committees on Bank Acts in the 1850s. Tooke was elected Fellow of the Royal Society in March 1821; shortly afterwards, with Ricardo, Malthus, James Mill and others, he founded the Political Economy Club and took a prominent part in its discussions until very late in his life. He died in London on

26 February 1858. A few days later, in a letter to Engels, Marx wrote: 'Friend Thomas Tooke has died, and with him the last English economist of any value' (Letter of 5 March 1858, in Marx and Engels 1983, p. 284).

Tooke's writings may be divided into two groups, two phases of his work which it is useful to distinguish for a better appraisal of his contribution. The first phase consists essentially in a systematic attempt to collect and analyse as much historical material and statistical information as possible, connected with price changes in England from 1793 onwards: a thorough observation of facts, aimed at understanding the determinants of fluctuations in the domestic price level. This phase is represented by his writings from 1823 to 1838 – from Tooke's first pamphlet *Thoughts and Details on the High and Low Prices of the Thirty Years from 1793 to 1822* to the first two volumes of his *History of Prices,* that is to say, with his *Considerations on the State of the Currency* (1826) *and* the two *Letters* to Lord Grenville (1829) in between. In the second group of writings, Tooke finally brings into focus and elaborates a few significant general principles which he gradually became certain could be derived from his observation of facts; moreover, he fully perceives the conflict between those principles and the prevailing notions, and copes with the arguments of his critics. This second phase of Tooke's work covers the writings from volume 3 of the *History* (1840) to Volumes 5 and 6 (1857, a year before Tooke's death) and comprises his *Inquiry into the Currency Principle* (1844) – the most representative and outstanding piece, together with volume 4 of the *History* (1848), of Tooke's voluminous work.

The main result emerging from Tooke's observation of facts in the first group of writings can be summarized as follows: the great fluctuations of prices that occurred in the 45 years following 1792 must be attributed to circumstances affecting the conditions of supply of commodities, rather than to the alterations in the system of the currency – the latter being represented by the suspension of convertibility from 1793 and its resumption after 1819 (by the Resumption Act of 1819). The prevailing view was that the value of the currency had been depreciated by the suspension, and enhanced by the contraction in the amount of circulating medium that the resumption of convertibility was alleged to have brought about. According to Tooke, 'the most extensive induction of facts' made it apparent that the phenomena of high prices from 1792 to 1819 and of the comparatively low prices after 1819, did not originate in the variations in the quantity of money (independently of whether the latter proceeded from the alterations in the system of the currency or from any other cause). The great fluctuations of prices originated instead from alterations in the cost of production and from other 'accidents' affecting supply: the character of the seasons (more unfavourable on the average from 1793 to 1818 than from 1818 to 1837); marked variations in the cost of imported commodities, as well as in the existence and removal of various obstacles (revolutions, wars) from the several sources of foreign supply; significant improvements in machinery and sciences generally, all tending to reduce the cost of production of numerous commodities (or to provide cheaper substitutes). In volume 2 of the *History*, the rate of interest is listed for the first time amongst the causes of the high and of the low prices in the period under consideration – 'a higher rate of interest constituting an increased cost of production' and 'a reduction of the general rate of interest' leading 'to reproduction at a diminished cost' (1838, pp. 847 and 849). As we shall see, this is in our view the crucial point upon which hinge those aspects of Tooke's contribution that are most relevant for the modern scholar of capitalism.

The connection between money and prices occupies the centre of the stage in the second group of Tooke's writings: 'the prepossession or prejudice', as he puts it, that the quantity of money must have a direct influence on the prices of commodities. By 'money' must be understood, Tooke insists in pointing out to the supporters of the Currency Theory, not only coin and paper money (banknotes), but also cheques, bills of exchange, settlements and whatever form of paper credit which may come to be a component part of the circulating medium, performing the

functions of money in daily transactions. By 1844 he was fully convinced

> that the prices of commodities do not depend upon the quantity of money indicated by the amount of bank notes, nor upon the amount of the whole of the circulating medium; but that, on the contrary, the amount of the circulating medium is the consequence of prices. (1844, p. 123)

Tooke's evidence for this conclusion was ultimately the fact that the banks, including the Bank of England, did not appear to have the power to add to the quantity of money in circulation – unless other independent circumstances, such as an extension of trade and a rise in prices, were 'coincidently' in progress (1844, p. 66); nor did the banks appear to have the power to diminish the total amount of the circulation. Banks may withhold loans and discounts, and may refuse any longer to issue their own notes, but those loans, discounts and notes will be replaced in due course, Tooke argues, 'by other expedients calculated to answer the same purpose' (1844, p. 122). Only compulsory paper money issued directly by a government in payment for goods and services (like the French *assignats*) constitutes a fresh source of demand, so that alterations in its quantity act directly as an originating cause on prices (see 1844, pp. 68–78; 1848, pp. 183–97).

The power of the banks to expand and contract the quantity of the circulating medium at pleasure, was taken for granted by the Currency School and by most writers. It was a challenging task, for Tooke and the Banking School, to convince those writers of the lack of such a power, and, in consequence, of the fact that such alterations in the quantity of money as do actually occur are the *effect* of increased transactions and prices, and not the *cause* of them. Tooke has the great merit of having succeeded in bringing into focus the heart of the matter: the question of the effects of changes in the rate of interest on the inducement to purchase commodities. 'Abundance of money' – that is, a high disposition on the part of the banks to make advances in the way of loan or discount – results in the first place in a high price of securities and a low rate of interest; thus the power of the banks to add to the amount of the circulating medium, and hence to act as an originating cause on trade and prices, will ultimately depend on whether a low rate of interest supplies the *stimulus* to purchase commodities. Tooke points out that actual experience does not validate the notion that the facility of borrowing at a low rate of interest, not only confers the power of purchasing commodities, but also affords the motive and inducement to do it. 'The error', he says, 'is in supposing the *disposition* or *will* to be co-extensive with the power' (1844, p. 79). No relation of cause and effect between variations in the rate of interest and variations in the demand for commodities can be inferred from trustworthy evidence (compare 1857, vol. 5, p. 345).

The questions of the connection between money and prices and between the rate of interest and the price level are thus clearly seen as two sides of the same coin. Arguing against the dominant opinion that a low rate of interest raises prices and that a high rate depresses them, Tooke actually maintained that a persistent reduction in the rate of interest constitutes a reduction in the cost of production, which could not fail, by the competition of the producers, to bring about a fall of prices (compare 1844, p. 81). He went so far as to state that it is difficult to find evidence of facts more in contrast with the influence ascribed to a low rate of interest in raising prices and vice versa: 'The theory is not only not true, but the reverse of the truth' (1844, p. 84).

It is important to notice that Tooke's conception of the relation between the rate of interest and the price level, and the connected notion of 'endogenous money' (as we would now call his view of the relation between money and prices), are in no way contingent upon the particular currency system of his day, with the relevant part played in it by precious metals. Rather, it is the denial of any power on the part of the banks to regulate at will the amount of the circulating medium, together with the emphasis on the circumstances affecting supply, that provide Tooke with the basis for his criticism of the idea that every influx or efflux of the precious metals must cause a rise and fall of prices, independently of circumstances connected with the cost of production of commodities. On that same basis he opposes the prevailing view that the discovery of a

gold mine within the premises of the Bank of England – Ricardo's famous assumption in his first pamphlet (1811) – would necessarily raise the prices of commodities. And he argues that for an increased production of gold to be associated with a permanent rise in the prices of commodities, measured in gold, the increased production must be the consequence either of the discovery of more fertile mines, or of improved methods of working the existing ones (1848, pp. 199 ff.; 1857, vol. 6, pp. 413–4).

Monetary policy questions, naturally, permeate all of Tooke's writings. The chief place amongst them is occupied by the Bank Charter Act of 1844 and the controversies that both preceded and followed its implementation – controversies centred upon the idea of a separation of the business of the note issue from the banking business of the Bank of England. This idea, opposed by Tooke, was given statutory effect by the Act, and brought about in due time many of the shortcomings that had been foreseen by Tooke (for an extensive critical account of Tooke's views on banking policy, see Gregory 1928, 1929, vol. 1).

On several policy issues that are still relevant today, the modern orthodox scholar of monetary questions and central banking policy is likely to find himself more in agreement with Tooke's views than with those of the supporters of the Currency School. This hardly applies, however, to those views of Tooke's which are more strictly connected with his conception of the relation between money and prices, and of the influence of the rate of interest on the price level. As an important example of one such view, one may refer to Tooke's contention that, as the Bank of England and the banks collectively cannot arbitrarily change the amount of the circulating medium, nor operate through that medium on the prices of commodities, the only 'infallible means' they have to influence foreign exchanges – 'so as to arrest a drain, or to resist an excessive influx' – is by a forcible operation on securities: a great advance in the rate of interest on the one hand, or a great reduction of it in the other. Now, the articulated line of argument laid down by Tooke in discussing the power of the central bank to influence foreign exchanges (cf. 1844, pp. 123–4), appears on the whole no less alien from today's quantity of money approach to problems of general prices than it was in Tooke's day.

Besides Marx (see above, and also 1857–8; 1859), Tooke's most outstanding contemporaries who praised his work and ideas were Malthus (1823) and J.S. Mill (1844, 1852, ch. 24, pp. 203–4). (Malthus's appreciation, however, must not be overrated: he tends to understand Tooke's early contribution merely as a confirmation of his own views on value against those of Ricardo – namely, 'that everything must be attributed to supply and demand', rather than simply to 'labour and the costs of production'; 1823, p. 218.) The most strenuous opponent of Tooke's ideas and policy recommendations was Robert Torrens (1840, 1844, 1848). This author's criticisms grew increasingly severe as Tooke's work advanced with the development of more general principles from the empirical analyses. By 1844, Tooke's thesis that the prices of commodities do not depend upon the quantity of money is referred to and criticized as 'the most astonishing of the many astonishing fallacies' (Torrens 1844, p. 43). Torrens's criticisms are extensively dealt with by Tooke in volume 4 of the *History of Prices* (1848), and by Fullarton (1844) and Wilson (1847).

If Torrens was the most outstanding critic of Tooke amongst classical economists, Knut Wicksell, the father of 20th-century monetary theory, has been his most outstanding critic since the inception of marginalism. Wicksell's conceptions ultimately constitute the main reference-point of this century's (not so large) literature in which Tooke's work and ideas are somehow taken into consideration, starting from Gregory's *Introduction* to the *History of Prices* (1928). In fact, we can look today at Tooke and Wicksell as the chief exponents of two alternative ways of reasoning about the connection between money and prices. Wicksell's criticisms of Tooke's view are somewhat vitiated by their being mostly based upon the interest elasticity of the demand for loan capital, as postulated by the marginalist theory (see 1898, ch. 7; 1906, pp. 175–208). There is, however, one

important criticism which does not reflect Wicksell's tendency to superimpose upon Tooke's view his own theory. He criticizes Tooke's reasoning about the effect of the rate of interest on the cost of production and commodities prices, as entailing that every persistent move in either direction would cause a progressive divergence of both interest and prices from their initial levels: a persistent reduction in the rate of interest

> would lead to a reduction ... in the demand for loans by business people, money would flow into the banks and would cause a further reduction of interest rates, and so on, until the rate fell to nil – In other words, the money rate of interest would be in a state of unstable equilibrium. (Wicksell 1906, p. 187.

This conclusion actually follows, not from Tooke's view of the influence of the rate of interest on prices, but from his conception of the rate of interest as a magnitude 'entirely governed by the supply of and demand for monied capital', on which the central bank can exercise only a *temporary* influence (1826, sect. I; 1857, pp. 556–7; see also Newmarch 1857, pp. 66–72).

Not to have acknowledged that the monetary authorities do have the power of determining the rate of interest – albeit a power exercised under a wide range of constraints – constitutes, in our opinion, the main shortcoming of Thomas Tooke and the Banking School.

Selected Works

1823. *Thoughts and details on the high and low prices of the last thirty years from 1793 to 1822*. London: John Murray.

1826. *Considerations on the state of the currency*. London: John Murray.

1829a. *A letter to Lord Grenville, on the effects ascribed to the resumption of cash payments on the value of the currency*. London: John Murray.

1829b. *A second letter to Lord Grenville, on the currency in connexion with the corn trade and on the corn laws, to which is added a postscript on the present commercial stagnation*. London: John Murray.

1838. *A history of prices and of the state of the circulation from 1793 to 1837*, 2 vols. London: Longman, Orme, Brown, Green & Longmans.

1840. *A history of prices and of the state of the circulation in 1838 and 1839, with remarks on the corn laws and some of the alterations in our banking system*. London: Longman, Orme, Brown, Green & Longmans.

1844. *An inquiry into the currency principle; the connection of the currency with prices and the expediency of a separation of issue from banking*, Series of reprints of scarce works on political economy no. 15, 2nd ed. London: London School of Economics and Political Sciences, 1959.

1848. *A history of prices and of the state of the currency from 1839 to 1847 inclusive: With a general review of the currency question and remarks on the operation of the Act 7 & 8 Vict. 32*. London: Longman, Brown, Green & Longmans.

1856. *On the Bank Charter Act of 1844, its principles and operations; with suggestions for an improved administration of the Bank of England*. London: Longman, Brown, Green & Longmans.

1857. (With W. Newmarch.) *A history of prices and of the state of the circulation during the nine years 1848–1856* (in two volumes; forming the fifth and sixth volumes of the history of prices from 1792 to the present time). London: Longman, Brown, Green, Longmans & Roberts.

Bibliography

Fullarton, J. 1844. *On the regulation of currencies; being an examination of the principles on which it is proposed to restrict, within certain fixed limits, the future issues on credit of the Bank of England, and of the other banking establishments throughout the country*. London: John Murray.

Gregory, T.E. 1928. *An introduction to Tooke and Newmarch's A history of prices and of the state of the circulation from 1792 to 1856*, Series of reprints of scarce works on political economy no. 16. London: London School of Economics and Political Sciences, 1962.

Gregory, T.E. 1929. *British banking statutes and reports, 1832–1928,* 2 vols. London: Oxford University Press.
Johnson, A. 1856. *Currency principles versus banking principles; being strictures on Mr Tooke's pamphlet on the Bank Charter Act of 1844*. London: Richardson Brothers.
Malthus, T.R. 1823. Review of Tooke's 'Thoughts and details on the high and low prices of the last thirty years'. *Quarterly Review* 29: 214–239.
Marx, K. 1857–8. *Grundrisse* [Rough Draft]: *Foundations of the critique of political economy.* Harmondsworth: Penguin, 1973.
Marx, K. 1859. *A contribution to the critique of political economy*. Moscow: Progress Publishers, 1978.
Marx, K., and F. Engels. 1983. *Collected works*, vol. 40. London: Lawrence & Wishart.
Mill, J.S. 1844. Review of Tooke's 'An inquiry into the currency principle'. *Westminster Review,* March–June.
Mill, J.S. 1852. *Principles of political economy, with some of their applications to social philosophy*, 3rd ed. London: John W. Parker and Son.
Newmarch, W. 1857. Evidence before the select committee of the House of Commons on Bank Acts, 5th June. In Gregory (1929, vol. 2).
Ricardo, D. 1811a. The high price of bullion, a proof of the depreciation of bank notes. In *The works and correspondence of David Ricardo*, vol. 3, 4th ed, ed. P. Sraffa. Cambridge: Cambridge University Press, 1966.
Ricardo, D. 1811b. Reply to Mr Bosanquet's practical observations on the Report of the Bullion Committee. In *The works and correspondence of David Ricardo*, vol. 3, ed. P. Sraffa. Cambridge: Cambridge University Press, 1966.
Torrens, R. 1840. *A letter to Thomas Tooke, esq. in reply to his objections against the separation of the business of the Bank into a department of issue, and a department of deposit and discount: With a plan of bank reform*. London: Longman, Orme, Brown, Green & Longmans.
Torrens, R. 1844. *An inquiry into the practical working of the proposed arrangements for the renewal of the Charter of the Bank of England, and the regulation of the currency: With a refutation of the fallacies advanced by Mr Tooke*. London: Smith, Elder & Co.
Torrens, R. 1848. *The principles and practical operation of Sir Robert Peel's Bill of 1844 explained and defended against the objections of Tooke, Fullarton, and Wilson*. London: Longman, Brown, Green & Longmans.
Wicksell, K. 1898. *Interest and prices*. London: Macmillan, 1936.
Wicksell, K. 1906. *Lectures on political economy.* Vol. 2: *Money.* London: Routledge & Kegan Paul, 1962.
Wilson, J. 1847. *Capital, currency, and banking; being a collection of a series of articles published in the 'Economist' in 1845, on the principles of the Bank Act of 1844, and in 1847, on the recent monetarial and commercial crisis; concluding with a plan for a secure and economical currency.* London: *The Economist.*

Torrens, Robert (1780–1864)

B. A. Corry

Keywords

Absolute advantage; Bank Charter Act (1844); Banking School; Capital theory of value; Colonization; Comparative advantage; Convertibility; Corn-ratio theory of profits; Currency School; Invariable standard of value; Labour theory of value; Real bills doctrine; Reciprocal tariffs; Ricardo, D.; Sraffa, P.; Terms of trade; Torrens, R.

JEL Classifications

B31

Torrens, if not in the top rank of the classical economists, or in the class for example of Ricardo, Senior or John Stuart Mill, certainly was of the second rank and the equal of, or even above, James Mill or McCulloch in terms of originality, theoretical reasoning and the range of economic topics that he considered. His work was almost completely neglected in the years after his death in 1864 and his re-emergence to his rightful place as an important member of the Classical School was initially due to Seligman in his famous article 'On Some Neglected British Economists' (1903) and later to the definitive study by Lionel Robbins (1958). In recent years Torrens has also come to the fore again because of the debates surrounding the Sraffa interpretation of Ricardo.

Robert Torrens was a most prolific writer and produced a vast quantity of books and pamphlets on all sorts of economic matters for over 50 years. His first publication appeared in 1808 (*The Economists Refuted*) and his last in 1858 (*Lord Overstone on Metal and Paper Currency*). He managed all of this against the background of an extremely busy life that included several different careers. He was a professional soldier – a colonel in the Marines – and was decorated for gallantry at the battle of Anholt. Subsequently he became the

proprietor of the *Globe* newspaper, a Member of Parliament, the planning genius behind the colonization and development of New South Wales, a founder member of the Political Economy Club and many other things besides. He even found time to write two never-read novels, the *Hermit of Killarney* and *Coelibia in Search of Husband*, both of which contain hefty chunks of economic discourse.

His specific contributions to economics may be dealt with under the general headings of microeconomics, theory of money and banking, commercial policy and colonization.

Torrens's main contributions to microeconomics concerned the Ricardian system. He objected to the search for an absolute, invariant measure of value and also tried to replace the labour quantity theory of value with a capital theory of value – where relative commodity values are determined by relative capital inputs. He did not however fully realize that his definition of capital combining wages and materials was different from Ricardo's which was really only wages.

On the other hand, and somewhat inconsistently, it is now clear (see Langer 1982; De Vivo 1985) that Torrens fully understood the corn-ratio theory of profits that Sraffa ascribed to Ricardo, and that he (Torrens) derived this from Ricardo. He also saw, following Ricardo that given the agricultural rate of profit, the price of manufactured goods relative to corn, was given. All of this is clearly spelt out in the second edition of *An Essay on the External Corn Trade* (1820) and must therefore lead one to doubt Hollander's argument (1979) that Ricardo did not mean the corn-ratio theory of profit and the key role of the agriculture sector in his analysis to be taken too seriously.

In the first edition (1815) of the *Essay* Torrens has a clear statement of the principle of comparative advantage well before its more popularly ascribed origins in Ricardo's *Principles* (1819). He makes a clear distinction between absolute and comparative advantage and indeed he actually hints at this distinction in his earlier *Economist Refuted* (1808).

In the field of money and banking Torrens is best known for his championing of the Currency School in their debate with the Banking School. Essentially the currency principle was that a mixed currency, that is, a currency consisting of notes and coins, should be regulated so that movements in it were the same as under a purely metallic currency. Unlike the bullionists, however, the Currency School did not believe that convertibility alone would achieve the conformability of a mixed currency to a metallic one. To this end, Torrens may claim to have been the originator of the plan, activated in the Bank Charter Act of 1844, to separate the issue and banking department of the Bank of England. This he did in his *Letter to Lord Melbourne* (1837) and he later vigorously defined the legislation in his *Principles and Practical Operation of Sir Robert Peel's Bill of 1844* (1848).

Students of Torrens find an inconsistency with this aspect of his monetary thinking and his earlier espousing of the anti-bullionist position; in particular his *Essay on Money and Paper Currency* (1812) is a strong plea for a paper currency without convertibility and relying on the real bills doctrine to prevent excess issue. The reasons for abandoning this extreme anti-bullionist position and his switch to the Ricardian line are explained in his *On the Means of Establishing a Cheap, Secure and Uniform Currency* (1828).

Torrens's main contribution to the theory of commercial policy was to suggest a modification of the general classical case for free trade. He pointed out, and was amongst the first to do so, that a country might alter its terms of trade in its favour by use of an import tariff. In a series of letters to Lord John Russell (published in 1844 as *The Budget*) he argued the case for what he termed 'reciprocity', that is, if some countries had tariffs unilateral free trade was a mistaken policy and in these cases reciprocal tariffs should be adopted. Against the change that he was abandoning the central classical (Ricardian) belief in free trade, Torrens replied that he was just applying the logic of the Ricardian analysis.

Finally, Torrens had a significant influence on the theory and practice of colonization. Along with most of the later classical economists he rejected the Smithian view that colonies were of no economic benefit to the colonial power. Much

of the later classical case for colonies was based on the view that colonies would provide profitable investment outlets to offset a declining rate of profit at home. Torrens used this argument in some of his later writings but his main argument was that colonies were an ideal solution to the Malthusian overpopulation problem. In this view he was undoubtedly influenced by his interpretation of the causes of Irish poverty – a country incidentally where Torrens was born and in general its problems had a profound effect on his thinking.

In terms of colonization policy Torrens, like Wakefield, was opposed to the movement of labour on to free land on the grounds that this would lead to a dispersed population and land holdings of suboptimal size. He advocated systematic colonization with the price of land set sufficiently high that large units of capital would have to be amassed before the immigrant labourers became independent farmers.

Selected Works

1820. *An essay on the external corn trade*. London: Longman, Rees, Orme, Brown & Green.
1821. *An essay on the production of wealth*. London: Longman, Hurst, Rees, Orme, & Brown.
1835. *Colonization of South Australia*. London: Longman, Rees, Orme, Brown, & Green.
1837. *A letter to the right honourable Lord Viscount Melbourne*. London: Longman, Rees, Orme, Brown & Green.
1844. *The budget*. London: Smith, Elder & Co.
1847. *On the operation of the Bank Charter Act of 1844*. London: Smith & Elder.

Bibliography

De Vivo, G. 1985. Robert Torrens and Ricardo's 'corn-ratio' theory of profits. *Cambridge Journal of Economics* 9: 89–92.
Langer, G.F. 1982. Further evidence for Sraffa's interpretation of Ricardo. *Cambridge Journal of Economics* 6: 397–400.
Robbins, L.C. 1958. *Robert Torrens and the evolution of classical economics*. London: Macmillan & Co. (This book includes brief summaries of all of Torrens's writings.)
Seligman, E.R.A. 1903. On some neglected British economists. *Economic Journal* 13: 511–535.

Total Factor Productivity

Diego Comin

Abstract

Total factor productivity (TFP) is the portion of output not explained by the amount of inputs used in production. This article sets out the measurement and importance of TFP for growth, fluctuations and development as well as likely future directions of research.

Keywords

Endogenous growth; Innovation; Patents; Real business cycles; Research and development; Solow residual; Technology; Total factor productivity

JEL Classifications

O4

Total factor productivity (TFP) is the portion of output not explained by the amount of inputs used in production. As such, its level is determined by how efficiently and intensely the inputs are utilized in production.

TFP growth is usually measured by the Solow residual since Solow (1957). Let g_Y denote the growth rate of aggregate output, g_K the growth rate of aggregate capital, g_L the growth rate of aggregate labour, and alpha the capital share. The Solow residual is then defined as $g_Y - \alpha^* g_K - (1 - \alpha)^* g_L$. The Solow residual accurately measures TFP growth if (a) the production function is Cobb–Douglas, (b) there is perfect competition in factor markets, and (c) the growth rates of output and the inputs are measured accurately.

TFP plays a critical role on economic fluctuations, economic growth and cross-country per capita income differences. At business cycle frequencies, TFP is strongly correlated with output and hours worked. Based on this observation, Kydland and Prescott (1982) initiated the real business cycle (RBC) literature. In the standard business cycle model, shocks to TFP are propagated by pro-cyclical labour supply and investment, thereby generating fluctuations in output and labour productivity at business cycle frequencies with an amplitude that resembles the US data. Subsequent work has introduced pro-cyclical fluctuations in measured TFP by incorporating unmeasured labour hoarding and/or capacity utilization in the standard framework (see, for example, Burnside et al. 1995; Basu 1996; King and Rebelo 1999). In this way, TFP fluctuations can be driven by shocks to aggregate demand in addition to the standard interpretation that attributes them to aggregate supply shocks.

As shown in the landmark article by Robert Solow (1956), long-run growth in income per capita in an economy with an aggregate neoclassical production function must be driven by growth in TFP. For over 30 years, the conceptual difficulty when trying to endogenize TFP growth was how to pay for the fixed costs of innovation in a perfectly competitive economy with constant returns to scale in capital and labour. In this context, all output is exhausted by paying capital and labour their marginal products; therefore, no resources are left to pay for the innovation costs. Romer (1990) and Aghion and Howitt (1992) solved this problem by granting the innovator monopolistic rights over his innovation, which are sustainable through the patent system. In this way, innovators can recoup the initial fixed costs of innovation through the profit margin they make from commercializing their patent.

By linking the TFP growth rate to innovation, endogenous growth models shed light on the determinants of TFP growth. R&D subsidies and an abundance of skilled labour reduce the marginal cost of conducting R&D and increase the rate of innovation development and, therefore, the TFP growth rate. Expanding markets increase the innovators' revenues, leading to more innovation and higher TFP growth.

Solow (1956) also demonstrated that cross-country differences in technology may generate important cross-country differences in income per capita. Klenow and Rodriguez-Clare (1997) and Hall and Jones (1999) have confirmed that most of the gap in income per capita between rich and poor countries is associated with large cross-country differences in TFP. Cross-country differences in TFP can be due to differences in the physical technology used by countries or in the efficiency with which technologies are used. To explore the relative importance of these factors, it is necessary to have data on direct measures of technology. Comin, Hobijn and Rovito (2006) put together direct measures of technology adoption for approximately 75 different technologies and show that the cross-country differences in technology are approximately four times larger than cross-country differences in income per capita. Further, technology is positively correlated to income per capita. Thus, cross-country variation in TFP is, to a large extent, determined by the cross-country variation in physical technology.

Likely Future Directions

Economic Fluctuations

Recognizing that a large portion of TFP growth is caused by endogenous innovation decisions has significant implications for the business cycle. This is likely to be an important research topic. Comin and Gertler (2006) show that low-persistence, non-technological shocks generate pro-cyclical fluctuations in the market value of innovations. Agents arbitrage these innovation opportunities and generate a procyclical rate of innovation development and, hence, of TFP growth. The model-induced fluctuations in TFP are as large and persistent as in the data. More important, by linking a component of TFP to innovation activity, TFP becomes a mechanism that propagates low-persistence shocks, thus increasing its persistence, rather than a source of disturbances as in standard RBC models. This same logic can be extended to other processes that determine the endogenous level of technology, such as endogenous technology

adoption processes, which are more relevant in developing economies. This may be an important ingredient to understanding high and medium-term fluctuations in developing economies.

Long-run Growth

A significant fraction of innovations are not patented. For some, this is because they are not embodied in any new good or are not a recipe for a new chemical process and, therefore, are not patentable. Others are not patented because innovators simply decide not to apply for a patent. Three important areas of research are to understand (a) how important patents are for innovation activity, (b) the determinants of nonpatentable innovations and (c) how they interact with the patentable R&D type of innovations that fit the properties of the Romer (1990) and Aghion and Howitt (1992) models.

Two papers have argued that patents are not necessary for the innovators to recoup the innovation costs. Innovators in Hellwig and Irmen (2001) can obtain rents to cover innovation costs despite being perfectly competitive because they face an increasing marginal cost of producing the intermediate goods that embody their innovations. Boldrin and Levine (2000) model innovation in perfectly competitive settings. In their model, to copy an innovation it is necessary to purchase one unit of the good that embodies it. Hence, the innovator is the monopolist of the first unit produced, and the revenues he extracts from selling it may cover the innovation costs, making up for a lack of patent protection.

Comin and Mulani (2006) model the development of disembodied innovations such as managerial and organizational techniques, personnel, accounting and work practices, and financial innovations. These are very different from embodied innovations in that the rents extracted by the innovators are not associated with selling the innovation per se. This has some interesting implications. First, the revenues accrued by the innovator–producer originate from the increased efficiency in producing his good or service with the innovation. If the innovator–producer has some monopolistic power in the market for his good or service, the increased efficiency from using the innovation in production yields an increase in profits that may cover the innovating costs. Second, since the innovator–producer's gain from innovating comes from the increased efficiency of production, the marginal private value of developing disembodied innovations is increasing in the value of the firm. This has important cross-sectional and time-series implications. In the cross-section, firms with higher values (resulting from larger sizes or ability to charge higher markups) have more incentives to develop disembodied innovations. In the time series, shocks that reduce the value of the firm reduce its incentives to develop disembodied innovations. One such shock may be an increase in the probability that a competitor steals the market. If the occurrence of this shock requires the development of a new patentable product, the model implies the possibility of an aggregate trade-off between investments in developing disembodied innovations and embodied innovations. A complete understanding of the determinants of these different types of innovation may be critical for explaining secular TFP dynamics.

Development

Understanding the determinants of technology adoption is key to explaining crosscountry variation in TFP. On the theory side, an increasing number of theories link the adoption of technologies to the role of institutions (Acemoglu et al. 2007), financial markets (Alfaro et al. 2006; Aghion et al. 2006), endowments (Caselli and Coleman 2006) and policies (Holmes and Schmitz 2001). The challenge is to bring these theories to the data and assess their empirical relevance. The new country-level data on measures of micro technologies must be an important input towards this goal.

See Also

- ▶ Endogenous Growth Theory
- ▶ Real Business Cycles
- ▶ Technology

Acknowledgment I thank Steven Durlauf for helpful comments.

Bibliography

Acemoglu, D., P. Antras, and E. Helpman. 2007. Contracts and technology adoption. *American Economic Review* 97: 916–943.

Aghion, P., and P. Howitt. 1992. A model of growth through creative destruction. *Econometrica* 60: 323–351.

Aghion, P., Comin, D., and P. Howitt. 2006. When does domestic saving matter for economic growth? Working Paper No. 12275. Cambridge, MA: NBER.

Alfaro, L., Chanda, A., Kalemi-Ozcan, S., and S. Sayek. 2006. How does foreign direct investment promote economic growth? Exploring the effects of financial markets on linkages. Working Paper No. 12522. Cambridge, MA: NBER.

Basu, S. 1996. Procyclical productivity: Increasing returns or cyclical utilization? *Quarterly Journal of Economics* 111: 719–751.

Boldrin, M., and D. Levine. 2000. *Growth under perfect competition*. Los Angeles: Mimeo, UCLA.

Burnside, C., M. Eichenbaum, and S. Rebelo. 1995. Capital utilization and returns to scale. In *NBER macroeconomics annual*, ed. B.S. Bernanke and J.J. Rotemberg. Cambridge, MA: MIT Press.

Caselli, F., and J. Coleman. 2006. The world technology frontier. *American Economic Review* 96: 499–522.

Comin, D., and M. Gertler. 2006. Medium term business cycles. *American Economic Review* 96: 523–551.

Comin, D., and S. Mulani. 2006. A theory of growth and volatility at the aggregate and firm level. Working Paper No. 11503. Cambridge, MA: NBER.

Comin, D., Hobijn, B., and E. Rovito. 2006. Five facts you need to know about technology diffusion. Working Paper No. 11928. Cambridge, MA: NBER.

Hall, R., and C. Jones. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114: 83–116.

Hellwig, M., and A. Irmen. 2001. Endogenous technical change in a competitive economy. *Journal of Economic Theory* 101: 1–39.

Holmes, T.J., and J.A. Schmitz Jr. 2001. A gain from trade: From unproductive to productive entrepreneurship. *Journal of Monetary Economics* 47: 417–446.

King, R., and S. Rebelo. 1999. Resuscitating real business cycles. In *Handbook of macroeconomics*, vol. 1B, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Klenow, P., and A. Rodriguez-Clare. 1997. The neoclassical revival in growth economics: Has it gone too far. In *NBER macroeconomics annual*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Romer, P. 1990. Endogenous technological change. *Journal of Political Economy* 98(5): S71–S102.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70(1): 65–94.

Solow, R. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39: 312–320.

Tournaments

Tor Eriksson

Abstract

Tournament theory is a theory of promotion-based incentives which also contributes to the understanding of firms' wage structures and individual earnings. In tournaments wage-rank differentials act as an incentive scheme when firms cannot directly observe employees' effort. Empirical evidence is mainly from sports and lab experiments while there are fewer studies of businesses and organizations.

Keywords

Pay; Promotions; Raises; Relative performance; Wage structures

JEL Classifications

M51; J31

A significant proportion of employees work together with other workers, some of whom are performing the same tasks. Moreover, for many jobs it is difficult to observe individual performance. Under such conditions relative performance schemes are frequently used a mechanism for motivating employees. These can take the form of competition for promotions or schemes that award tenure when performance exceeds a certain standard. Because of the similarity of these reward schemes to those found in professional sports competitions, scholars have named them tournaments. Tournament theory aims at explaining promotions and raises associated with

them. As a considerable share of pay increases occurs through job (title) changes, the theory complements other explanations of individual earnings differentials. Tournament models also contribute to our understanding of firms' wage structures and differences therein.

The Basic Model

The basic tournament model developed in Lazear and Rosen (1981) has risk-neutral agents, a firm with two identical contestants (employees), operating in a competitive industry. The employee's utility is a function of her net income, that is, the difference between income from work and the cost of effort. Output is the sum of the two employees' effort and two stochastic disturbances: a common and an individualspecific component. The firm does not directly observe the employees' efforts. The payment scheme is fixed in advance and has a winner's and a loser's prize to the better and poorer performer of the contestants, respectively. The prizes are independent of levels as well as of differences in employees' performance. The worker's problem is to choose her optimal level of effort taking into account the firm's compensation scheme. In other words she maximises her expected utility, which depends on the probability of winning the two prizes and her cost of effort.

As the contestants are assumed to be identical ex ante, the probability of winning is 0.5 and they choose the same equilibrium level of effort, which depends on two factors. The first is the difference between the winner's and the loser's prize. A larger prize spread induces employees to compete harder for earning the promotion and hence to exert more effort. Second, the more important individualspecific random components in output are, the less effort the contestants will provide.

The firm also needs to set wage levels high enough to attract workers to participate in the tournament. The solution is that the firm chooses a wage spread that induces employees to put forth effort up to the point where the marginal cost of effort equals the marginal benefit of it to the firm. Thus, the tournament compensation scheme is efficient and brings about first-best level of effort. Luck also affects the firm's decision. To maintain a given level of effort, an increase in the random component in output has to be offset by an increase in the wage spread.

Extensions and Predictions

The Lazear-Rosen article contained analyses where some of the assumptions of the basic tournament model were relaxed, and this line of research continued in a series of articles (Green and Stokey 1983; Nalebuff and Stiglitz 1983; O'Keeffe et al. 1984) which offered several extensions; see also the survey by McLaughlin (1988). These papers examined a number of additional variations on tournaments such as winning by a gap, rewards depending on distance from the loser's performance and multiple prizes. Empirically, these are of rather limited interest and have not been followed up in the recent literature.

Assumptions that turned out to affect equilibrium outcomes qualitatively were risk-neutrality, only two contestants, homogeneous agents and the one-stage tournament. The key result from relaxing the assumption of risk-neutrality is that the optimal level of effort is lower than the first-best level as risk-averse contestants want income insurance. Thus, risk aversion narrows the wage spread which yields lower effort. Does the size of the tournament matter? Intuitively, the larger the number of contestants, the lower the expected probability of winning, so in order to induce same level of effort, the firm has to widen the wage spread. Consequently, the spread increases in tournament size. With risk-averse agents this result is fragile (McLaughlin 1988), but for relevant tournament ranges the spread is plausibly increasing.

Allowance for heterogeneous contestants has been modelled by assuming that ability differentials equal differences in marginal cost of effort. Two information structures for heterogeneity in ability been studied: (i) agents know their own but not their contestant's ability, and (ii) full knowledge. In both cases mixing high and low ability agents gives rise to inefficiency; in (i) because of no sorting and in (ii) because the tournament is not attractive to low performers. In the asymmetric

information case, more knowledge (such as entry credentials) is needed. Alternatively a larger pay spread can induce self-sorting. In the full knowledge case, segregating low and high performers also leads to inefficiency (in both leagues). Handicaps may overcome some of the problems: typically at the expense of efficiency, however.

In many firms there are usually several rounds of promotion competitions over workers' careers. Rosen (1986) analyses the multistage tournament, more precisely a sequential, elimination tournament like in professional tennis. In this setting promoted employees earn not only a raise but also the expected value of continued competition. With a constant spread, effort would decline through rounds (as the tournament shrinks in size) and would be lowest at the top. Rosen shows that in order to keep effort unchanged, the wage spread has to grow linearly with rank, and will make a jump at the top to compensate for the absence of further competition. An implication of the analysis is that the firm uses a convex pay-rank schedule as a means for providing incentives to all its employees.

A key prediction of tournament theory is that increased wage spread yields higher effort and output. However, the participation constraint puts an upper bound on the optimal wage spread. When cooperation among employees is important, or when jobs are strongly interdependent, relative performance games may give rise to too strong incentives for uncooperative behaviour. Extending the basic model to allow for employees to behave strategically against their rivals shows that sabotaging behaviour lowers equilibrium effort (Lazear 1989). Thus, firms might also compress their internal wage structures on efficiency grounds. Another form of strategic employee behaviour that can lower effort is collusion among employees. This (and sabotaging) is more likely when there are few contestants. Mitigating harmful effects by increasing the tournament size is costly. As the contestant pool grows, it becomes more heterogeneous giving rise to increased inefficiency. Tournament models also predict that firms will favour insiders over outsiders (Chan 1996).

Several papers compare (the efficiency of) tournaments with other individual incentive pay schemes, in particular piece rates. The question that has not been addressed much is why do (especially larger) firms use tournament structures? The principal advantage of tournaments is that contestants are insulated from common risks. Thus, tournaments are predicted to be more common in risky environments. The other main advantage is that it is less costly to measure relative than absolute performance. A third, but less well understood, possibility is that promotions also can serve other functions, notably sorting of employees.

Empirical Tests

Tournament theory yields quite a number of testable predictions. These have been tested on data from sports, laboratory experiments and businesses. The studies examine whether observed empirical patterns are consistent with tournament models, but none has to my knowledge tested tournaments against the other theoretical explanation for promotions: promotions as signals (Waldman 1984).

Following Ehrenberg and Bognanno's (1990) study of the effects of prize structures on professional golfers' performance, a large empirical sports literature has built up, providing in the main evidence in support of several of tournament theory's predictions. Sports data are rich and of high quality, but to what extent the results are generalizable to firms remains an open question.

Almost from the beginning tournaments have been studied by means of lab experiments that allow researchers to control for a host of parameters while examining a number of treatments, and also to compare with other incentive schemes. Beginning with Bull et al. (1987), several experimental studies have confirmed the predictions about the impact of wage spread on performance but have found a high variance in subjects' effort. In general many of the other predictions like those concerning tournament size, number of rounds and degree of uncertainty have been supported by the laboratory evidence. Again, generalizability to businesses is an issue. Some key concepts like cost of effort or attitudes to risk are very

difficult to observe outside the lab. However, as has been shown in some recent experiments, endogenous sorting can have profound effects on results. Generalizing from the lab to the upper echelons of organizations can in particular be associated with external validity problems.

Evidence outside sports and the lab is rather scarce. The evidence that exists is often based on data from a single firm, specific industries or firms' managerial personnel. Many studies have tested the prediction that a larger wage spread increases employees' effort and hence firm performance. The hypothesis finds support in many, but not all, studies. Another implication that has been tested is the convexity of firms' wage structures. These results are, however, consistent not only with tournament theory but also with other theories. A problem in testing tournaments is that there are potential alternative explanations for individual predictions. Thus, a positive relation between pay spread and firm performance, and convex within-firm wage structures, are also consistent with convex returns to ability due to magnification effects (in hierarchies bosses' decisions matter more) and with the promotions-as-signals hypothesis. More direct tests concern the hypotheses of a positive relation between number of contestants and the magnitude of the raise from promotion, and that tournament organizations favour insiders. One way to arrive at more convincing evidence is to test several hypotheses on the same data set. Only a few studies (e.g. Eriksson 1999; Knoeber and Thurman 1994) have done this. A number of single-firm studies have documented more facts about wage and promotion dynamics within firms. Although one should be cautious in treating them as stylized facts, a fruitful avenue for further research on tournaments is to extend tournament models to account for patterns observed in these studies.

See Also

- ▶ Hierarchy
- ▶ Incentive Compatibility
- ▶ Payment Systems
- ▶ Sliding Scales (Wages)

Bibliography

Bull, C., A. Schotter, and K. Weigelt. 1987. Tournaments and piece-rates: An experimental study. *Journal of Political Economy* 95: 1–33.

Chan, W. 1996. External recruitment versus internal promotion. *Journal of Labor Economics* 14: 555–570.

Ehrenberg, R., and M. Bognanno. 1990. Do tournaments have incentive effects? *Journal of Political Economy* 98: 1307–1324.

Eriksson, T. 1999. Executive compensation and tournament theory: Empirical tests on Danish data. *Journal of Labor Economics* 17: 262–280.

Green, J., and N. Stokey. 1983. A comparison of tournaments and contracts. *Journal of Political Economy* 91: 349–364.

Knoeber, C., and W. Thurman. 1994. Testing the theory of tournaments: An empirical analysis of broiler production. *Journal of Labor Economics* 12: 155–179.

Lazear, E. 1989. Pay equality and industrial politics. *Journal of Political Economy* 97: 561–580.

Lazear, E., and S. Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89: 841–864.

McLaughlin, K. 1988. Aspects of tournament models: A survey. In *Research in labor economics*, vol. 9, ed. R. Ehrenberg, 225–256. Greenwich: JAI Press.

Nalebuff, B., and J. Stiglitz. 1983. Prizes and incentives: Towards a general theory of compensation and competition. *Bell Journal of Economics* 14: 21–43.

O'Keeffe, M., K. Viscusi, and R. Zeckhauser. 1984. Economic contests: Comparative reward schemes. *Journal of Labor Economics* 2: 27–56.

Rosen, S. 1986. Prizes and incentives in elimination tournaments. *American Economic Review* 76: 701–715.

Waldman, M. 1984. Job assignments, signalling, and efficiency. *Rand Journal of Economics* 15: 255–267.

Townshend, Hugh (1890–1974)

Victoria Chick

Keywords

Liquidity preference; Loanable funds; Townshend, H.; Walras's law

JEL Classifications

B31

Townshend is an anomaly amongst economists, for he owes his reputation almost entirely to one brilliant article. He took a First in mathematics at Cambridge in 1912 and stayed on to prepare for the civil service examinations under Keynes's supervision. He served with the Post Office, where his duties included economic forecasting.

His correspondence with Keynes over the just-published *General Theory* (Keynes 1979) reveals both the extent of Townshend's intellectual grasp of that complex and difficult book and something, whether derived from his studies with Keynes or innate in his temperament, which allowed him to accept aspects of *The General Theory* which others resisted. These qualities bore fruit in the famous 12 page article, published as a note in the *Economic Journal* (1937a). This note takes issue with Hicks's attempt, in his review of the *General Theory* (Hicks 1936; Keynes 1936), to transform the theory of liquidity preference into a mirror image of loanable funds theory by Walras's Law. Townshend saw that this was an attempt to retain the link between prices and the flow concepts of cost and demand. In contrast, he argued, it was in the nature of Keynes's liquidity preference theory that expectations of the future could change the value of assets overnight and be reflected in market prices of those assets even in the absence of actual trading. Thus current prices could be determined by subjective as well as objective factors and future prices were indeterminate.

Townshend's achievement was to 'follow liquidity preference theory where it led: to the destruction of determinate price' (Shackle 1967). Keynes had left this implicit.

Townshend's restatement required the courage to go against established modes of thought: whether by temperament or because they are imbued with outdated conceptions of science, most economists are determinists. Townshend's stock can only rise as the methodological change that has occurred in science becomes known across the divide between science and the arts.

Townshend also wrote four book reviews for the *Economic Journal* (1937b, 1938, 1939, 1940) which show a breadth of conception and keenness of intellect from which one wishes economics had benefited more.

Selected Works

1937a. Liquidity-premium and the theory of value. *Economic Journal* 47: 157–169.
1937b. Review of R.G. Hawtrey, *Capital and employment. Economic Journal* 47: 321–326.
1938. Review of G.L.S. Shackle, *Expectations, investment and income*. *Economic Journal* 48: 520–523.
1939. Review of H. Munro, *Principles of monetary-industrial stability. Economic Journal* 49: 102–105.
1940. Review of F. Hayek, *Profits, interests and investment. Economic Journal* 50: 99–103.

Bibliography

Hicks, J.R. 1936. Mr. Keynes's theory of employment (review article). *Economic Journal* 46: 238–253.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.
Keynes, J.M. 1973, 1979. *Collected writings of J.M. Keynes*, vols. 14 and 29, ed. D.E. Moggridge. London: Macmillan.
Shackle, G.L.S. 1967. *The years of high theory: Invention and tradition in economic thought, 1926–1939*. Cambridge: Cambridge University Press.

Toynbee, Arnold (1852–1883)

A. Kadish

Keywords
Economic history; Industrial revolution; Liberalism; New liberalism; Toynbee, A

JEL Classifications
B31

Arnold Toynbee, best known for his lectures on the Industrial Revolution (published posthumously in May 1884), was during the late 1870s and early 1880s a major influence on the shape and direction of the interest at Oxford in socio-economic questions and their history.

Born in London on 23 August 1852, Arnold Toynbee was the fourth child and second son of Dr. Joseph Toynbee, FRS, a philanthropist and a successful aural surgeon. Initial plans for his education at Rugby were first delayed by an accident at the age of 13 or 14 which resulted in severe concussion and, in the long run, in recurring migraines, impeding prolonged mental exertion. These plans were finally shelved for financial reasons following Joseph Toynbee's death in a laboratory accident. After two unprofitable years in a military preparatory school, and some classes at King's College London, Toynbee's education was reduced to long periods of solitary reading.

He developed an independent if unsystematic bent, coupled with overconfidence in his capability to master on his own any subject which might catch his attention.

Having come into a modest inheritance from his father's estate at the age of 21, Toynbee entered Pembroke College, Oxford, in January 1873 with the intention of reading for Greats. He shortly afterwards migrated to Balliol, which he found socially and intellectually more attractive. For health reasons he eventually settled for a Pass degree, obtained in 1878. However, he had impressed the college sufficiently to be offered a tutorship in charge of the candidates for the Indian civil service. In 1881 he was appointed Senior Bursar, and at the time of his death was about to be elected Fellow. Toynbee had meanwhile become involved in a number of public causes, all of which may be seen as related to efforts to revive liberalism as a radical-reformist movement. These included Church reform aimed at the democratization of Church of England government on the parish level, adult education through the cooperative movement, rural reform, Irish land reform, and municipal politics. In 1883 he stood, unsuccessfully, as a liberal candidate for one of the north ward seats on the Oxford City Council (represented by T.H. Green until his death in 1882) and Toynbee may well have contemplated the possibility of a political career. His academic interests reflect his search for a scientifically reasoned programme for comprehensive reform, while his inquiries were greatly influenced by a fear of the consequences of rising working-class radicalism.

The Industrial Revolution lectures, delivered in 1881–2, offered a liberal interpretation of industrialization and its political, social and economic consequences as an alternative to both socialist and laissez-faire views of industrial society. Despite their ideological bias and fragmentary form they constituted an important departure in English economic history. They demonstrated the usefulness of an historical approach to the study of industrial society, thereby suggesting an alternative to economic theory, which at the time was increasingly regarded as either morally or ideologically unacceptable or as inapplicable to current conditions. Hence Toynbee directly contributed to the Oxford inclination towards an empirical and historical approach to the study of socio-economic questions. In addition, Toynbee outlined the possibility of an historical and, thereby, a relativist consideration of economic theories, seen as reflecting historical circumstances in which they were formed and, while of a limited general application, of considerable interest to the historian. Finally, and perhaps most importantly, the Industrial Revolution lectures suggested an autonomous approach to the study of economic history, based on its own type of primary sources, in which economic circumstances were not placed in causal subservience to political developments (as in the work of W. Cunningham). Toynbee's approach was ideologically and philosophically acceptable to the next generation of economic historians. It was not materialistic yet it tended to regard economic change in terms of general impersonal trends rather than attributing it to the conscious action of narrow interest groups (as in the work of J.E.T. Rogers).

Following an attempt to convince a hostile London audience of the futility of land nationalization and the desirability and attainability of liberal alternatives, Toynbee suffered a nervous breakdown. While convalescing he contracted meningitis and died at the age of 30, widely hailed as a martyr to the cause of social harmony and to the type of reformism which became known as New Liberalism. His reputation and martyrdom contributed to the popularity of the university settlements, beginning with Toynbee Hall, and a general interest at Oxford and Cambridge in social and economic reform. His importance as an economic historian,

however, emerged only with the development of the study some years after his death.

See Also

► Industrial Revolution

Selected Works

A partial collection of Toynbee's work including a reconstruction of his Industrial Revolution course based on his and on some of his students' notes was edited by his friend Alfred Milner and published posthumously as *Lectures on the industrial revolution in England. Popular addresses, notes and other fragments,* London: Rivingtons, 1884. *'Progress and poverty': A criticism of Mr. H. George: Being two lectures delivered in . . . London* (London: Kegan Paul, 1883) was published in pamphlet form and appended to the 2nd edition (1887) of *Lectures on the industrial revolution in England*.

Bibliography

Kadish, A. 1986. *Apostle Arnold. The life and death of Arnold Toynbee (1852–1883)*. Durham: Duke University Press.

Montague, F.C. 1889. *Arnold Toynbee*. Baltimore: Johns Hopkins University Studies in Historical and Political Science.

Tozer, John Edward (1806–1877)

G. Campanelli

Keywords

Fixed capital; Mathematics and economics; Tozer, J. E.; Whewell, W.

JEL Classifications

B31

Tozer was born at Woolwich in 1806 and died in London in 1877. He was privately educated and at the age of 26 was admitted as a 'Pensioner' of Caius College, Cambridge. He immediately showed his ability in mathematics by carrying off the first prize in 1833 and 1834. In the four years after his graduation in 1836, Tozer wrote two essays on mathematical economics, the first on 'Machinery' (1838), and the second on 'Landlords' (1840). However, soon after the publication of these two papers he abandoned economics and went into law, in which he achieved distinction. Later he went into university administration.

Tozer's two papers on economics represent a systematic application of mathematical reasoning to political economy. In that period Whewell, at Trinity College, Cambridge, was trying to introduce mathematical analysis into economics and Tozer adopted his method. Like Whewell, Tozer believed that mathematics, because it turned economics into a 'science' characterized by a series of propositions leading to 'axiomatic truths', was not only appropriate but necessary to the subject (Tozer, 1838, pp. 1–2).

Of the two essays on economics, that on 'Machinery' is by far the more interesting. In this paper, Tozer wants to provide a mathematical basis for the idea that the employment of machinery always increases the wealth of community. His final conclusion is that the capitalist is 'not only . . . unable to secure his own advantage at the expense of any other class, he cannot even prevent a general participation in the benefit' (Tozer, 1838, p. 10).

However, the most original feature of this paper is Tozer's mathematical treatment of the problem of machinery. In particular the calculation of the annuity and the algebraic formulation of the construction period of machinery can undoubtedly be regarded as a sophisticated contribution to the fixed-capital debate of that time.

See Also

► Machinery Question

Selected Works

1838. *Mathematical investigation of the effect of machinery on the wealth of a community in which it is employed and on the fund for the payment of wages.* Cambridge: Cambridge Philosophical Society, Transaction 6.
1840. *On the effect of the non-residence of landlords on the wealth of a community.* Cambridge: Cambridge Philosophical Society, Transaction 7.

Tradable and Non-tradable Commodities

A. D. Woodland

Abstract

The distinction between internationally tradable and non-tradable commodities lies at the heart of the reason for the development of the theory of international trade as an area of economics distinct from the general theory of value. The existence of nontradable commodities as well as tradable commodities implies that some markets are domestic while others are international. Connections between the prices of these two sets of commodities have been the subject of the famous factor-price equalization and Stopler–Samueslon theorems. Non-tradable goods play a prominent role in the analysis of many problems, such as tariff reform, exchange rates and international transfers.

Keywords

Bastable, C. F.; Complementarities; Factor-price equalization; Globalization; Heckscher–-Ohlin–Vanek model; Interest rate differentials; International portfolio Ccoice; International trade theory; Mill, J. S.; Project evaluation; Purchasing power parity; Real exchange rates; Reduced form analysis; Ricardo, D.; Rybczynski theorem; Shadow pricing; Sticky prices; Stolper–Samuelson theorem; Substitutes and complements; Tariffs; Terms of trade; Torrens, R.; Tradable and non-tradable commodities

JEL Classifications

F1

The distinction between internationally tradable and non-tradable commodities lies at the heart of the reason for the development of the theory of international trade as an area of economics distinct from the general theory of value. If the theory of value approach were adopted, nations would simply be a collection of production and consumption units, each with its own monetary and political system. The international trade literature has, however, imposed the assumption that certain classes of commodities are non-tradable. Accordingly, there exist purely domestic markets as well as international markets, and this is arguably the most important distinguishing feature of the international trade literature.

The early classical economists made a clear distinction between products which were assumed to be internationally tradable, and factors of production such as land, labour and capital, which were assumed to be non-tradable internationally but perfectly mobile within each nation. This distinction, which is central to the writings of such eminent classical theorists as Ricardo, Torrens, J.S. Mill and Bastable, was carried on by early twentieth-century writers such as Taussig (1927), Yntema (1932), Ohlin (1933), Haberler (1936), and Mosak (1944). The modern literature, initiated perhaps by Samuelson (1953), has continued with the same basic model structure.

The traditional international trade model, therefore, contains non-tradable commodities. However, they are in the form of non-producible factors of production rather than products, and this provides important structure to the traditional model. Moreover, and perhaps more importantly, further structure is typically imposed by assuming that the non-tradable factors are in fixed supply. As a result, consumer preferences do not play a

role in the market for non-tradable commodities. It therefore follows that the traditional model of trade encompasses non-tradable commodities in a very special way.

A more general, and more useful, way to deal with non-tradable commodities is to define a set of commodities that are non-tradable, and allow this set to include products as well as factors. While the existence and importance of non-tradable commodities in the form of products was recognized as early as, for example, Ohlin (1933, ch. 8), Haberler (1936, pp. 34–5) and Taussig (1927, ch. 5), it was not until much later that non-tradable products were explicitly incorporated into international trade models. For earlier surveys of this literature, see McDougall (1970) and Woodland (1982, ch. 8).

There is the fundamental question of why certain commodities are not traded internationally. Some may be non-traded because of their intrinsic nature; they are simply not transportable. Others may be transportable and hence tradable but are not traded because it is unprofitable to do so due to the costs of transportation or other expenses such as tariffs. Finally, products may be tradable, but trade in them may be illegal – an extreme form of trade quota.

While it has long been recognized that there is a difference between a commodity being tradable and being traded, and that the difference arises as a result of the profitability of trade, few models actually deal with non-traded commodities in this way. Rather, it is typically assumed that transport costs are zero for some commodities (tradable) and infinite, or, at least, sufficiently large for others (nontradable) so as to preclude their trade. Exceptions do exist. Hadley and Kemp (1966) and Woodland (1968) explicitly model transportation and thus endogenize the division into traded and non-traded commodities. Xu (2003), using a continuum of goods with transport costs and tariffs, has the boundary between traded and non-traded goods endogenous. Melitz (2003) models a continuum of firms with different levels of productivity producing a continuum of varieties in which only the more productive firms export, the remainder producing varieties only for the domestic market.

Relationships Between Domestic and International Markets

The existence of non-tradable commodities as well as tradable commodities implies that some markets are domestic while others are international. However, this does not mean that the domestic markets operate in isolation from international markets. On the contrary, the prices of domestic (non-tradable) commodities will be influenced by activities in the international markets. Though of less apparent interest, the reverse is also true: the prices of internationally traded commodities are influenced by activities in domestic markets. The various national markets for non-tradable commodities are, of course, connected only indirectly via the international markets for tradable commodities.

Let $p = (p_t, p_n)$ denote the partition of the price vector for a nation into tradable and non-tradable commodities, and let $X(p) = (X_t(p), X_n(p))$ denote the vector of excess supply functions correspondingly partitioned. If the foreign nation's functions and variables are distinguished by an *, the perfectly competitive equilibrium conditions for internationally tradable commodities and the non-tradable commodities of the home and foreign nations may be written as

$$X_t(p_t, p_n) + X_t^*(p_t, p_n^*) = 0 \tag{1}$$

$$X_n(p_t, p_n) = 0 \tag{2}$$

$$X_n^*(p_t, p_n^*) = 0 \tag{3}$$

Under reasonable regularity conditions the market equilibrium conditions for the domestic commodities in the home nation, (2), may be solved for p_n as a function of p_t as $p_n = P_n(p_t)$. Similarly, (3) may be solved as $p_n^* = P_n^*(p_t)$. Thus, the equilibrium conditions for domestic commodities provide the connection between the prices of tradable commodities and the prices of domestic or non-tradable commodities. An elegantly simple analysis of the connection between the markets for tradable and non-tradable commodities is provided by Jones (1974).

A significant amount of the international trade literature has been devoted to the relationship between prices of tradable and non-tradable commodities. Within the context of factors being the only non-tradable commodities, there are several famous propositions that emerge. The Stolper–Samuelson (1941) theorem indicates, within a two-product, two-factor model, that if the relative price of one product increases then one factor price will increase proportionately more and the other factor price will fall. The factor whose price increases, and whose real income therefore increases, is the one used relatively intensively by the product whose price increases. Jones (2006) provides a historical perspective on this theorem and the literature it generated. Second, the factor-price equalization theorem provides conditions under which the factor-price vectors, here p_n and p^*_n, are the same in the two nations despite differences in factor endowments. The sufficient conditions are that each nation should have the same production technology and that their endowment vectors be sufficiently close so that each nation is diversified and produces the same set of traded goods (Samuelson 1953; McKenzie 1955; Dixit and Norman 1980). Recent contributions include Blackorby et al. (1993) and Chakrabarti (2006), who provide necessary and sufficiency conditions.

The Role of Non-tradable Commodities

The reason non-tradable commodities complicate the analysis of many problems in international trade theory is that, whenever there is a disturbance to equilibrium, there will be an effect on the markets for non-tradable goods. The price of non-tradable goods will have to adjust to restore equilibrium. This adjustment of prices will, in general, affect the variable of interest and may yield different qualitative results than are obtained from a model without non-tradable commodities (factors or products). Much of the literature is concerned with the question of how the introduction of non-tradable products affects results obtained from models with only tradable products. However, some recent literature has gone the other way, enquiring whether trade in one or more factors alters results obtained on the assumption that all factors are non-tradable.

As an example of the role that non-tradable products can play, consider an increase in the international price for a small open economy's import good. If there are just two traded products, a single consumer, and all non-tradable commodities are factors in fixed supply, then it is well known that the quantity of imports will fall. The introduction of a third product that is produced and consumed but not traded internationally can upset this result if there are sufficient complementarities in production or consumption. The rise in the price of the imported product causes imports to change directly, and indirectly via the consequent change in the price of the non-tradable product. At the initial price of the non-tradable good, the import price increase induces higher home production and lower demand since the income effect is unambiguously negative. This direct effect implies lower imports, just as when non-tradable commodities don't exist. However, if imports and the non-tradable good are net complements, an excess supply of the non-tradable good ensues from the import price increase (ruling out inferiority). Its price then falls to clear the domestic market. This fall in price causes a reduction in the net supplies of both the imported and the non-tradable goods since they are net complements. If this indirect reduction in the net supply of the importable good is sufficiently strong to outweigh the direct positive effect of the increase in its price, the quantity of imports will rise. Thus arises the paradoxical case where an increase in the price of the imported product causes the level of imports to rise. This case occurs because of the assumed net complementarity between the imported and non-tradable products.

It is noteworthy that many of the difficulties that occur when non-tradable products are included in a model arise in the consumption sector. In the example of the previous paragraph, if there is a fixed demand for the non-tradable product the indirect income effect vanishes and so the quantity of imports falls in response to a rise in their price. Alternatively, one can ensure this result by assuming that the nontradable and imported products are net substitutes.

For some problems the existence of non-tradable commodities has no substantial influence upon the formal solution. The prices of non-tradable products can be eliminated from the equilibrium conditions by first solving for their prices in terms of the prices for tradable goods, and then substituting into the excess supply functions for tradable goods. This yields the excess supply functions expressed in terms of tradable goods' prices, and the equilibrium conditions reduce to

$$\tilde{X}(p_t) = X_t(p_t, P_n(p_t)) = 0 \qquad (4)$$

These resulting 'reduced form' excess supply functions can then be used directly to analyse various problems. For example, the usual stability conditions for international equilibrium can be applied to the reduced form excess supply functions on the assumption that domestic markets clear instantly. The problem of the effect of a transfer of income upon the terms of trade can be similarly handled. For details on the reduced form approach to non-tradable goods see Dixit and Norman (1980, pp. 89–92) and Woodland (1982, pp. 172–3, 218–22). Those that prefer to deal with the structural form include Komiya (1967), McDougall (1970), and Jones (1974).

Situations Where the Role of Non-tradable Commodities Is Prominent

For other problems, or where interest centres on the variables in the structural model relating to non-traded goods, the existence of non-tradable commodities has to be explicitly taken into account. Some specific instances are as follows:

1. In the case of shadow pricing of commodities for the purpose of evaluating the welfare effects of a public project in the presence of tariffs, the existence of non-tradable commodities provides special complications. While tradable commodities should be evaluated using world prices, the appropriate shadow prices for non-tradable commodities depend in a complex way upon technology and taste conditions (Warr 1982; Dinwiddy and Teal 1987).
2. Several analyses of technological advances in the production of a traded product upon the output levels of other traded goods, and upon the real exchange rate, have focused attention upon the role played by non-tradable commodities. In general, the effect of the 'boom' upon the other tradable good's production is ambiguous, stemming partly from the adjustments in the market for non-tradable commodities. For details, see Corden and Neary (1982) and references therein.
3. The effect of the introduction of non-tradable products upon the Stolper–Samuelson, Rybczynski and factor-price equalization theorems was thoroughly analysed by Ethier (1972). In the case of the Stolper–Samuelson theorem, a change in the relative price of the non-tradable products induces a change in the price of the non-tradable product. This has a further effect upon factor prices. The question of whether a particular factor gains in real income depends upon the directions of these price changes, and can be answered only from knowledge of the technology and preferences. In the case of the factor-price equalization theorem, Mainwaring (1978) demonstrates that non-traded goods may result in the non-equalization of national interest rates, while Deardorff and Courant (1990) argue that non-traded goods reduce the likelihood of factor-price equalization by reducing the size of the diversification cone.
4. The 'purchasing power parity' theory of exchange rates states that the relative value of currencies equals the relative purchasing power of each currency in its domestic markets. Clearly, the existence of non-tradable commodities with different prices across countries will cause the purchasing power of each currency to be different independently of the value of the exchange rate.
5. The responses to currency devaluation are also potentially affected by the existence of non-tradable products. An example of such an analysis is Dornbusch (1973), who concludes that his basic results hold up when a nontradable product is introduced into the model. Neary (1980) models sticky prices in domestic labour

and product markets and shows that short-run policy (for example, devaluation) responses are affected. More generally, non-traded goods affect macroeconomic variables in dynamic models. They explain persistent deviations from purchasing power parity and interest rate differentials (Backus and Smith 1993) and affect the responses to an oil price shock (Marion 1984), while productivity shocks in traded and non-traded goods sectors have different effects (Murphy 1986).

6. The welfare implications of tariff reform depend upon the existence of nontradable products. Following early work by McDougall (1970) and Dornbusch (1974) on the role of non-traded goods for tariff reform, Hatta (1977) and Fukushima (1979) have shown that the policy of reduction of the highest rate of import duty to the next highest is welfare improving if non-tradable and tradable products are net substitutes. If they are sufficiently complementary, the policy may reduce welfare. Non-traded goods are incorporated, and play important roles, in the analyses by Diewert et al. (1989, 1991) of tariff reform and in the trade restrictiveness index of Anderson et al. (1995). Xu (2003) shows that increased wage inequality can arise from the shifting boundary between traded and non-traded goods as a result of a tariff reform.
7. The effect of a transfer of income from abroad upon the welfare of a small open economy is affected by adjustments in the non-traded goods markets and may lead to the transfer paradox whereby the recipient's welfare declines, as shown by Yano and Nugent (1999). Schweinberger (2002) emphasizes the role of complementarities between non-traded and traded goods in this context.
8. Non-traded goods play a prominent role in the modelling of international portfolio choice and explanations for home country bias (Tesar 1993; Baxter et al. 1998).
9. Davis and Weinstein (2001) show that, while the Heckscher–Ohlin-Vanek model is empirically rejected, a model that accounts for non-traded goods is consistent with the data for trade by ten OECD countries.

Concluding Comments

It is somewhat surprising that non-tradable products (as opposed to factors) have only relatively recently attracted explicit attention in the literature, given that most economic activity is in the markets for domestic products. Perhaps more future studies will introduce non-tradable products automatically, except where a reduced form analysis is applicable, and give more attention to the endogeneity of the division of commodities into traded and non-traded groupings. Globalization, whereby the costs of undertaking international trade are being reduced by technological improvements in transport and communications and trade policy liberalization, will continue to change the boundary between traded and non-traded commodities. The trend towards more outsourcing of services and intermediate inputs from foreign countries provides a clear example.

See Also

- ▶ Factor Prices in General Equilibrium
- ▶ Globalization
- ▶ Tariffs
- ▶ Terms of Trade
- ▶ Trade Costs

Bibliography

Anderson, J.E., G.I. Bannister, and J.P. Neary. 1995. Domestic distortions and international trade. *International Economic Review* 36: 139–157.

Backus, D.K., and G.W. Smith. 1993. Consumption and real exchange rates in dynamic economies with non–traded goods. *Journal of International Economics* 35: 297–316.

Baxter, M., U.J. Jermann, and R.G. King. 1998. Nontraded goods, nontraded factors, and international non-diversification. *Journal of International Economics* 44: 211–229.

Blackorby, C., W. Schworm, and A. Venables. 1993. Necessary and sufficient conditions for factor price equalization. *Review of Economic Studies* 60: 413–434.

Chakrabarti, A. 2006. Factor price equalization beyond a 'cubic' world. *Economic Theory* 27: 483–491.

Corden, W.M., and J.P. Neary. 1982. Booming sector and deindustrialization in a small open economy. *Economic Journal* 92: 825–848.

Davis, D.R., and D.E. Weinstein. 2001. An account of global factor trade. *American Economic Review* 91: 1423–1453.

Deardorff, A.V., and P.N. Courant. 1990. On the likelihood of factor price equalization with nontraded goods. *International Economic Review* 31: 589–596.

Diewert, W.E., A.H. Turunen-Red, and A.D. Woodland. 1989. Productivity-and Pareto-improving changes in taxes and tariffs. *Review of Economic Studies* 56: 199–215.

Diewert, W.E., A.H. Turunen-Red, and A.D. Woodland. 1991. Tariff reform in a small open multi-household economy with domestic distortions and nontraded goods. *International Economic Review* 32: 937–957.

Dinwiddy, C., and F. Teal. 1987. Shadow prices for non-traded goods in a tax-distorted economy: Formulas and values. *Journal of Public Economics* 33: 207–221.

Dixit, A.K., and V. Norman. 1980. *Theory of international trade*. Cambridge: Cambridge University Press.

Dornbusch, R. 1973. Devaluation, money, and nontraded goods. *American Economic Review* 63: 871–880.

Dornbusch, R. 1974. Tariffs and nontraded goods. *Journal of International Economics* 4: 177–185.

Ethier, W. 1972. Nontraded goods and the Heckscher–Ohlin model. *International Economic Review* 13: 132–147.

Fukushima, T. 1979. Tariff structure, non-traded goods and theory of piecemeal policy recommendations. *International Economic Review* 20: 427–435.

Haberler, G. 1936. *The theory of international trade with its applications to commercial policy.* Trans. from the 1933 German edition by A. Stonier and F. Benham. London: William Hodge.

Hadley, G., and M.C. Kemp. 1966. Equilibrium and efficiency in international trade. *Metroeconomica* 18 (2): 125–141.

Hatta, T. 1977. A recommendation for a better tariff structure. *Econometrica* 45: 1859–1869.

Jones, R.W. 1974. Trade with non-traded goods: The anatomy of inter-connected markets. *Economica* 41: 121–138.

Jones, R.W. 2006. Protection and real wages: The history of an idea. *The Japanese Economic Review* 57: 457–466.

Komiya, R. 1967. Non-traded goods and the pure theory of international trade. *International Economic Review* 8 (2): 132–152.

Mainwaring, L. 1978. Interest rate equalization theorem with nontraded goods. *Journal of International Economics* 8: 11–19.

Marion, N.P. 1984. Nontraded goods, oil price increases and the current account. *Journal of International Economics* 16: 29–44.

McDougall, I.A. 1970. Non-traded commodities and the pure theory of international trade. In *Studies in international economics*, ed. I.A. McDougall and R.H. Snape. Amsterdam: North-Holland.

McKenzie, L.W. 1955. Equality of factor prices in world trade. *Econometrica* 23: 239–257.

Melitz, M.J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71: 1695–1725.

Mosak, J.L. 1944. *General equilibrium theory in international trade*. Bloomington: Principia Press.

Murphy, R.G. 1986. Productivity shocks, non-traded goods and optimal capital accumulation. *European Economic Review* 30: 1081–1095.

Neary, J.P. 1980. Nontraded goods and the balance of trade in a neo-Keynesian temporary equilibrium. *Quarterly Journal of Economics* 95: 403–429.

Ohlin, B. 1933. *Interregional and international trade*. Cambridge, MA: Harvard University Press. Reprinted1952.

Samuelson, P.A. 1953. Prices of factors and goods in general equilibrium. *Review of Economic Studies* 21: 1–20.

Schweinberger, A.G. 2002. Foreign aid, tariffs and non-traded private or public goods. *Journal of Development Economics* 69: 255–275.

Stolper, W., and P.A. Samuelson. 1941. Protection and real wages. *Review of Economic Studies* 9: 58–73.

Taussig, F.W. 1927. *International trade*, 1966. New York: Kelley.

Tesar, L.L. 1993. International risk-sharing and non-traded goods. *Journal of International Economics* 35: 69–89.

Warr, P.G. 1982. Shadow pricing rules for non–traded commodities. *Oxford Economic Papers* 34: 305–325.

Woodland, A.D. 1968. Transportation in international trade. *Metroeconomica* 20 (2): 130–135.

Woodland, A.D. 1982. *International trade and resource allocation*. Amsterdam: North-Holland.

Xu, B. 2003. Trade liberalization, wage inequality, and endogenously determined non-traded goods. *Journal of International Economics* 60: 417–431.

Yano, M., and J.B. Nugent. 1999. Aid, nontraded goods, and the transfer paradox in small countries. *American Economic Review* 89: 431–449.

Yntema, T.O. 1932. *A mathematical reformulation of the general theory of international trade*. Chicago: University of Chicago Press.

Trade and Environmental Regulations

Jenny Minier

Abstract

To what extent do environmental regulations affect trade flows? Early cross-sectional studies of the relationship generally found little

effect. More recent research that controls for industry heterogeneity, the endogeneity of trade policy, and the type of trade and industry has found sizeable, statistically significant effects of environmental regulations on trade. There is little evidence, however, that production of 'dirty' goods has shifted to countries with lax environmental standards.

Keywords

Environmental regulations; Footloose industries; Pollution haven; Trade policy

JEL Classifications

F18

The relationship between trade and environmental regulations has received much attention from trade negotiators, businesspeople, and environmental activists as well as academics. The intuition is straightforward: as a country increases an industry's environmental regulations (or the stringency with which they are enforced), the industry's costs increase, making the industry less competitive internationally, and increasing imports and decreasing exports in the industry. Ultimately, this could lead to the concentration of polluting industries in 'pollution havens', or countries with more lax environmental standards.

This conventional wisdom that environmental regulations affect trade flows – known in the academic literature as the 'pollution haven effect' – failed to receive much support in early papers on the subject, such as Grossman and Krueger (1993). A key result of this literature was that measures of environmental costs often had no effect (or even had positive effects) on a country's net exports. Evidence for the pollution haven effect is typically found by analysing the correlation between environmental costs and net trade flows. Because environmental standards are difficult to measure directly, and because there can be significant differences between standards as written and as enforced, many papers in the literature have used pollution abatement costs as a percentage of total costs for the measure of environmental standards. The basic hypothesis is that, if stringent environmental regulations are a source of competitive disadvantage, then a country's most regulated industries should have the highest levels of import penetration. Thus, early empirical studies typically involved cross-sectional regressions of industry trade flows on a set of industry characteristics, including environmental costs, and uncovered little evidence of a pollution haven effect. This was surprising, but was usually rationalized as either because environmental costs are a very small fraction of most firms' total costs (so even if a pollution haven effect existed, it was very close to zero), or owing to the 'Porter effect', the idea that regulation leads to innovation and lower costs.

However, this conclusion has largely been overturned by a new wave of empirical research that allows for unobserved industry effects and the endogeneity of environmental regulations. A potential problem with the traditional approach is unobserved heterogeneity: unobserved industry characteristics may be correlated with both trade flows and the degree of regulatory stringency. A second problem is simultaneity, or the potential for trade flows and pollution regulations to be determined simultaneously. For example, as argued in Ederington and Minier (2003), legislators may tend to relax enforcement of environmental policy when an industry faces increased imports (for example, if international trade agreements prevent them from responding by adjusting trade policy), effectively using environmental policy as a secondary trade barrier. In that case, failing to account for the endogeneity of environmental regulations results in biased estimates of the effect of environmental policy on trade flows. Levinson and Taylor (2008) express these points clearly, and provide a theoretical model to show likely sources of bias in conventional measures of pollution-haven effects. Clearly, unobserved heterogeneity in the pollution haven regression can be solved by using more disaggregated data and by using panel data, which allows the inclusion of industry fixed effects to capture unobserved industry characteristics; the endogeneity of environmental regulation can be addressed by using instrumental variables. Ederington and Minier (2003) and Levinson and Taylor (2008) use these approaches, and find

statistically and economically significant evidence of pollution haven effects.

Evidence for significant pollution haven effects creates potential concerns for the design and structure of trade agreements. As currently structured, the GATT differs in its treatment of trade and environmental policy: WTO members negotiate directly over tariff concessions, and are required to respect the resulting binding tariff ceilings, but retain discretion in setting environmental policy as long as the standards adhere to certain GATT rules (such as the national treatment provision). However, to the extent that countries can relax environmental regulations on import-competing sectors as a means of reducing trade flows, they can undermine commitments previously made with respect to trade policy. These issues have led to a large and growing body of work on the design and structure of trade agreements when domestic policies can be used as a secondary means of protection.

Of course, whether environmental regulations are used as a secondary trade barrier also raises the question of the extent to which environmental regulations are affected by trade concerns. There is definitely anecdotal evidence for such a connection: the earliest national environmental legislation (including, for example, the US Federal Water Pollution Control Act of 1970) required studies of the effects of environmental regulations on US competitiveness. In addition, several presidents have established task forces explicitly to relax domestic regulations (including environmental regulations) that adversely affected US trade competitiveness. However, explicit empirical evidence on this connection is limited (although Ederington and Minier 2003 do find that an increase in imports is correlated with a subsequent relaxation in pollution abatement costs).

It should be noted that this literature generally focuses on finding evidence for or against a pollution haven effect, or evidence that a tightening of environmental regulations deters exports (or stimulates imports). This is related to, but distinct from, the 'pollution haven hypothesis', that as trade barriers are reduced, pollution-intensive industries shift production into low-income countries with lax environmental regulations. As discussed in several papers, including Antweiler et al. (2001), evidence for a pollution haven effect does not necessarily imply the validity of the pollution haven hypothesis; while these effects are important, there are many other determinants (such as factor abundance) of trade flows. Antweiler et al. (2001) was one of the first papers to look for evidence of the pollution haven hypothesis, basically by linking trade liberalization to changes in pollution concentrations across countries. Ederington et al. (2004) subsequently attempted a more direct approach using data on trade flows: has trade liberalization induced the United States to produce cleaner goods while importing dirtier goods from abroad? Neither paper finds evidence for the pollution haven hypothesis. Similarly, Copeland and Taylor (2003) present results implying that free trade shifts production of 'dirty' goods toward capital-intensive, higher-income countries with more stringent environmental regulation.

Finally, the papers discussed to this point all examine the overall (cross-country, or cross-industry) relationship between trade flows and environmental regulations. Ederington et al. (2005) take a different approach, discussing several reasons why the effect of environmental policy on trade flows might be very important for some types of trade and some industries, although this effect could be masked in a full sample of industries. First, they show that differences in environmental standards are likely to matter most for trade between developed and developing economies (where standards are further apart), while the majority of trade occurs between developed countries. Second, they show that for most industries, environmental costs are a small part of total costs, so even a significant increase in environmental costs is unlikely to have much effect on production. Finally, they note that industries vary in the extent to which they are 'footloose' (that is, able to easily move production overseas to take advantage of lower environmental standards). Industries are considered to be less footloose if they have high transport costs (such as cement), high firm fixed costs, or seem to benefit from agglomeration economies (being geographically concentrated). Ederington et al. (2005) show that

the effect of environmental regulation on trade flows is large in magnitude and statistically significant for trade between developed and developing countries, for industries in which environmental costs are a large part of total costs, and for industries that are not footloose.

To summarize, the relationship between trade and environmental regulations has been much studied in the last two decades: it is a policy-relevant question that offered an interesting economic 'puzzle'. If environmental regulations make domestic firms less competitive internationally, why do we not see this effect in trade flows? Recent research that controls for the effects of policy endogeneity, industry heterogeneity, and the type of trade and industry suggests that environmental regulations do affect trade flows, and the magnitude of the effect is sizeable. There is little evidence, however, that production of dirty goods has shifted into countries with less stringent standards.

See Also

- ▶ Economic Development and the Environment
- ▶ Non-tariff Barriers
- ▶ Pollution Haven Hypothesis

Bibliography

Antweiler, W., B. Copeland, and M.S. Taylor. 2001. Is free trade good for the environment? *American Economic Review* 91: 877–908.

Copeland, B., and M.S. Taylor. 2003. *Trade and the environment*. Princeton: Princeton University Press.

Ederington, J., and J. Minier. 2003. Is environmental policy a secondary trade barrier? An empirical analysis. *Canadian Journal of Economics* 36: 137–154.

Ederington, J., A. Levinson, and J. Minier. 2004. Trade liberalization and pollution havens. *Advances in Economic Analysis and Policy* 4: 1–22.

Ederington, J., A. Levinson, and J. Minier. 2005. Footloose and pollution-free. *Review of Economics and Statistics* 87: 92–99.

Grossman, G.M., and A.B.. Krueger. 1993. Environmental impacts of a North American free trade agreement. In *The Mexico-U.S. Free trade agreement*, ed. P.M. Garber. Cambridge, MA: MIT Press.

Levinson, A., and M.S. Taylor. 2008. Unmasking the pollution haven effect. *International Economic Review* 49: 223–254.

Trade and Poverty

Nina Pavcnik

Abstract

This article reviews the relationship between trade and poverty, with specific focus on this link in developing countries. Since the 1980s, our understanding of the link between trade and poverty has been increasingly informed by empirical evidence. We first review the literature that emphasizes the dynamic effects of trade on poverty via growth. We next consider the literature on the static relationship between trade and poverty through changes in relative prices of goods and wages of the less educated. We conclude with a discussion of trade and child labour.

Keywords

Child labour; Foreign direct investment; Hecksher–ohlin trade theory; Import substitution; Labour mobility; Mercosur; Outsourcing; Poverty; Poverty alleviation; Returns to schooling; Skill-biased technical change; Stolper–samuelson theorem; Trade and poverty; Trade liberalization; Wage inequality

JEL Classifications

F

Trade liberalization is one of the most common policy prescriptions offered to initiate the process of poverty eradication in poor countries. Since the 1980s, our understanding of the link between trade and poverty in developing countries has been increasingly informed by empirical evidence. One part of the literature considers dynamic effects of trade on poverty via growth, while other parts emphasize the more static relationship between trade and poverty through changes in relative prices of goods and wages of the less educated.

Growth is potentially the most important channel through which trade might affect poverty. The usual argument goes as follows: trade promotes growth and growth leads to lower poverty. While many economists believe that these dynamic effects provide an important channel towards poverty reduction, the link between trade and poverty via growth has been empirically elusive. Many cross-country studies on trade and growth, perhaps best exemplified by Frankel and Romer (1999), find a positive association between trade and growth. But others, most notably Rodriguez and Rodrik (2001), have questioned the robustness of these findings and whether the positive association is indicative of good domestic policies and institutions in countries with economically sound trade policy. Growth could lower poverty by expanding employment and earnings opportunities of the poor, but growth could also bypass the poor (see Ravallion 2001). Two influential empirical studies by Dollar and Kraay (2002, 2004) conclude that growth is good for the poor by showing that average incomes of the poor move in tandem with average national incomes, and that trade via growth is good for the poor by showing that countries with bigger tariff cuts on average observed greater declines in poverty. Yet these findings remain a topic of heated academic debate, outlined in Deaton (2005) and Ravallion (2001).

Much of the academic debate on trade and poverty has focused on the static relationship between trade and poverty through changes in relative prices and wages of less educated workers. The stylized version of the Hecksher–Ohlin model predicts that trade should benefit the poor in a less developed country that is relatively well endowed with less educated labour and has a comparative advantage in unskilled labour-intensive goods. According to the Stolper–Samuelson theorem, trade liberalization will increase the return to the factor of production that is relatively abundant in a country and decrease the return to the scarce factor. Consequently, trade liberalization should reduce inequality between educated and less educated workers and lift the poor out of poverty in developing countries. In an influential project on trade and employment in developing countries, Krueger (1983) used this insight to argue in favour of outward-oriented trade regimes over import-substitution strategies adopted by many poor countries at that time.

Many developing countries liberalized trade during the 1980s and 1990s. As detailed micro-surveys of workers, households, and firms became more readily available, the researchers began to empirically examine the consequences of these reforms. Motivated by the academic debate on whether trade with poor countries in part contributed towards growing wage inequality in the OECD countries during the 1980s, these studies initially focused on the consequences of trade for wage inequality between more and less educated workers rather than for poverty. As reviewed in Wood (1999) and Goldberg and Pavcnik (2007), trade reforms were associated with increases (rather than the expected decreases) in inequality. Studies reviewed in Goldberg and Pavcnik (2007) concluded that the increases in the relative earnings of educated workers in poor countries could have been in part caused by an increase in the relative demand for educated workers in the aftermath of trade reforms due to trade-induced skill-biased technological change, outsourcing, or the higher skill intensity of exporting firms relative to non-exporters. But the studies remained silent on how trade reforms affected those at the bottom of the income distribution through economy-wide changes in the *absolute* demand for less skilled workers and through other channels ranging from consumption, industry wages, and unemployment to compliance with labour market standards. Only recently have studies focused on these issues.

Porto (2006) embeds heterogeneous households in a general equilibrium model of trade, where trade-induced relative price changes influence household welfare through changes in labour income (via the usual Hecksher–Ohlin mechanisms) and consumption. He explicitly accounts for the fact that households at the bottom of the welfare distribution are relatively less endowed with educated labour *and* spend a higher share of their household budget on basic items such as food than richer households. The model, combined with the estimates of key parameters from micro surveys, is used to simulate the effects of

Mercosur-induced changes in prices on welfare and poverty in Argentina. Although poor Argentine households were worse off after Mercosur because they tended to consume a relatively high share of the now more expensive unskilled-labour-intensive goods, these negative consumption effects were substantially smaller than the trade-induced increases in labour income of less educated workers.

A series of papers in the *Globalization and Poverty* project organized by Harrison (2007) consider the link between trade and poverty in a setting where labour is not perfectly mobile across industries and/or regions within a country. These studies examine whether individuals living in areas or industries that were more exposed to trade experienced smaller or bigger changes in poverty than less exposed areas or industries. The Ricardo–Viner model with industry-specific labour would predict that industry tariff cuts are associated with declines in relative industry wages. Since tariff declines in many developing countries were concentrated in sectors with lower wages and higher shares of unskilled workers before the reforms, this channel could in the short run increase poverty in areas with a higher pre-reform concentration of protected industries relative to the national trend. A study by Topalova (2007) shows that individuals living in rural Indian districts, where prereform employment was heavily concentrated in industries that experienced larger declines in tariffs, suffered increases in poverty relative to the national trade of declining poverty during the 1990s. Topalova conjectures that these individuals fare relatively worse because immobility, in part stemming from inflexible labour regulations, precludes them from reallocating from sectors that have been hit hardest by declines in tariffs toward sectors or areas that benefit from export expansion. A related study by Hanson (2007) finds that the poor living in Mexican states with higher concentration of export-oriented industries and inflows of foreign direct investment (FDI) fared better during the 1990s than the poor in areas with lower export and FDI exposure.

Overall, the studies of trade and poverty based on micro evidence suggest that there is substantial heterogeneity in how trade affects the poor within developing countries. Future work on trade and poverty needs to further examine barriers that inhibit the less educated from moving away from industries, areas, or firms that have been hit hardest by tariffs cuts to industries and firms benefiting from new exporting opportunities.

An important component of the policy debate on the effect of trade on the world's poor is the link between trade and child labour. Many are concerned that, because less developed countries are assumed to specialize in exports of low-skill products, high-income countries foster child labour in low-income countries by raising the demand for the products intensive in unskilled labour. In fact, as discussed in Edmonds and Pavcnik (2005a), the link between trade and child labour is far more complicated and depends on how trade affects family incomes and poverty, the availability of substitutes or complements for the child's work, the returns to education, and consumption prices in addition to how trade affects the demand for child labour. Empirically, the association between trade and living standards seems to be the dominant factor in how trade affects child labour. Cross-country evidence in Edmonds and Pavcnik (2006) provides no support for the claim that trade perpetuates high levels of child labour in poor countries via the demand channel. Similarly, Edmonds and Pavcnik (2005b) show that child labour declined in Vietnam following the rice market liberalization that relaxed rice export quota and improved the standard of living of many net-rice producing households, even though the employment opportunities in the rice sector increased. There are many reasons for a connection between child labour and family incomes or poverty. Ranjan (2001) in particular emphasizes the relaxation of credit constraints as important for understanding why growing trade might be associated with declining child labour despite rising demand for unskilled labour. In this vein, Edmonds et al. (2007) argue that, in the Indian context, the child's economic contribution to the household through the avoidance of schooling costs is important in understanding the interconnections among trade policy changes, child labour and schooling.

See Also

- ▶ Child Labour
- ▶ Heckscher–Ohlin Trade Theory
- ▶ Inequality (Measurement)
- ▶ Poverty
- ▶ Poverty Alleviation Programmes
- ▶ Trade Costs

Bibliography

Deaton, A. 2005. Measuring poverty in growing world (or measuring growth in a poor world). *Review of Economic Statistics* 87: 1–19.

Dollar, D., and A. Kraay. 2002. Growth is good for the poor. *Journal of Economic Growth* 7: 195–225.

Dollar, D., and A. Kraay. 2004. Trade, growth and poverty. *Economic Journal* 114: F22–F49.

Edmonds, E., and N. Pavcnik. 2005a. Child labor in the global economy. *Journal of Economic Perspectives* 18(1): 199–220.

Edmonds, E., and N. Pavcnik. 2005b. The effect of trade liberalization on child labor. *Journal of International Economics* 65: 401–441.

Edmonds, E., and N. Pavcnik. 2006. International trade and child labor: Cross-country evidence. *Journal of International Economics* 68: 115–140.

Edmonds, E., N. Pavcnik, and P. Topalova. 2007. *Trade adjustment and human capital investments: Evidence from Indian tariff reform*, Working paper No. 2884. Cambridge, MA: NBER.

Frankel, J.A., and D. Romer. 1999. Does trade cause growth? *American Economic Review* 89: 279–399.

Goldberg, P., and N. Pavcnik. 2007. Distributional effects of globalization in developing countries. *Journal of Economic Literature* 45: 39–82.

Hanson, G. 2007. Globalization, labor income, and poverty in Mexico. *Harrison* 2007.

Harrison, A., eds. 2007. *Globalization and poverty.* Chicago: University of Chicago Press and NBER.

Krueger, A. 1983. *Trade and employment in developing countries, vol. 3: Synthesis and conclusions*. Chicago: University of Chicago Press.

Porto, G. 2006. Using survey data to assess the distributional effects of trade policy. *Journal of International Economics* 70: 140–160.

Ranjan, P. 2001. Credit constraints and the phenomenon of child labor. *Journal of Development Economics* 64: 81–102.

Ravallion, M. 2001. Growth, inequality, and poverty: Looking beyond averages. *World Development* 29: 1803–1815.

Rodriguez, F., and D. Rodrik. 2001. Trade policy and economic growth: A skeptic's guide to cross-national evidence. In *NBER macro annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press for NBER.

Topalova, P. 2007. Trade liberalization, poverty, and inequality: Evidence from Indian districts. In Harrison (2007).

Wood, A. 1999. Openness and wage inequality in developing countries: The Latin American challenge to East Asian conventional wisdom. In *Market integration, regionalism and the global economy*, ed. R. Baldwin et al. Cambridge: Cambridge University Press.

Trade Costs

Jeffrey H. Bergstrand

Abstract

Trade costs refer to the additional costs paid potentially by the final consumer of a good or service beyond the price at which a producer sells a good. In international trade, such costs may include the transport cost from origin to destination, taxes (or tariffs) imposed by importing nations' governments, the costs of infrastructure to facilitate trade, the costs of communications, and foreign exchange costs.

Keywords

Communication costs; Currency unions; Distance; Economic geography; Exchange rate volatility; Factor content of trade; Foreign direct investment; Foreign exchange costs; Gravity equation; Information costs; Outsourcing; Portfolio flows; Regional trade agreements; Rent seeking; Sticky prices; Tariffs; Trade costs

JEL Classifications

F

'Trade costs' refer to the costs above and beyond the 'mill price' that the final consumer of a good (or service) pays. If a product is sold by producer i to consumer j at (mill) price p_i (in dollars), the consumer pays $p_i + T_{ij}$, where T_{ij} denotes the

'trade costs' (also in dollars). Such costs may cover the opportunity costs of resources, but some trade costs are just rent seeking barriers. International trade provides a fertile ground for studying the various types and economic effects of trade costs. International trade flows travel across large distances and empirical estimates of the costs of transporting goods between two economic centres are available. However, such flows also face less obvious trade costs. A broad interpretation of trade costs includes – beyond transport costs – information-gathering costs for a consumer to locate a foreign producer, financial and legal costs of negotiating contracts, policy-related barriers, and costs of final distribution in the importing country. As discussed in Anderson and van Wincoop (2004), the total trade costs associated with exporting a good from producer *i* to consumer *j* may be an average *ad valorem* add on of 170 per cent to the (mill) price of a good. In this article, we discuss some of the different types of international trade costs (subsets of which form national and local trade costs), and how such costs can influence the volume of trade between two nations and the relative prices of nations' goods.

Except for economic size, trade costs are probably the most important factor determining the volume of trade between a pair of countries (see Anderson 1979; Bergstrand 1985; and Anderson and van Wincoop 2003). Trade costs play a critical role in understanding outsourcing, the factor content of trade, the field of economic geography, foreign direct investment and foreign affiliate sales of multinational enterprises, and the proliferation of regional trade agreements in the post-war period. Obstfeld and Rogoff (2000) argue that trade costs in goods markets provide the critical common element that also explains at least *six major puzzles* in international macroeconomics.

The trade cost that comes immediately to mind is the cost of transporting a good from producer *i* to consumer *j*. EXW ('ex works') refers to the price of a good at the point of origin; mill price is a synonym for the ex works price ('mill' refers to where the good was produced). FOB ('free on board') refers to the price of a good delivered to and put 'on board' an overseas vessel. CIF ('cost, insurance, freight') refers to the price of a good to a named overseas port, including insurance costs. Empirical researchers have used both FOB export data and CIF import data but prefer the latter because import data measured at customs points is more accurate.

A common measure of international transport costs is consequently the difference between the CIF and FOB values of a trade flow. The International Monetary Fund (IMF) provides data on average 'CIF/FOB factors' $[100\times(\text{CIF value}-\text{FOB value})/\text{FOB value}]$ for countries. Baier and Bergstrand (2001) report that average CIF/FOB factors for 16 Organisation for Economic Co-operation and Development (OECD) countries in 1958 and 1988 were 8.2 per cent and 4.3 per cent, respectively. David Hummels (2001) finds that freight rates vary dramatically across countries with average transport costs ranging from 3.8 per cent of EXW price p_i for the United States to 13.3 per cent for land-locked Paraguay in 1994, varying even more across commodities within countries. Hummels (1999) finds evidence that inflation-adjusted tramp shipping rates have declined between 40 and 70 per cent from 1950 to 1995, but also finds evidence suggesting ocean shipping rates have not declined. It is common to express transport costs on an *ad valorem* (or rate) basis. Hence, the price faced by consumer *j* for producer *i*'s product, p_{ij}, can be expressed as $p_{ij} = p_i(1 + tc_{ij})$ where tc_{ij} is the (CIF − FOB)/FOB factor (for example, 0.04).

Another important trade cost is that associated with policy-related barriers imposed by national (or perhaps sub-national) governments. The trade cost most often envisioned here is a 'tariff', a tax imposed at customs points on imported goods. *Specific* tariffs are expressed in an amount of the home currency per unit imported good; T_{ij} used earlier denoted a specific trade cost. *Ad valorem* tariffs are expressed as a fraction of the value of the good; hence, an *ad valorem* tariff of ta_{ij} would cause the imported price of producer i's product for consumer *j* to be $p_{ij} = p_i(1 + ta_{ij})$ when the tariff is imposed on the FOB value and $p_{ij} = p_i(1 + tc_{ij})(1 + ta_{ij})$ when the tariff is imposed on

the CIF value. Other entries in this dictionary address tariffs in more detail.

Transport costs and tariff barriers are arguably the easiest trade costs to measure *directly*. Because of difficulty in measuring other types of trade costs directly, empirical economists have turned to *indirect* methods to estimate trade costs. Indirect methods fall into two basic categories: inferring trade costs from differences in trade volumes between pairs of countries and inferring trade costs from differences in prices between pairs of countries.

International trade economists have long used and increasingly applied the 'gravity equation' to explain empirically international trade flows (see Feenstra 2004, ch. 5; gravity equation). The gravity equation typically explains bilateral trade flows between country pairs using cross-sectional or panel data on pairs' gross domestic products (GDPs), bilateral distances between country pairs' economic centres, and several other variables representing bilateral trade costs to infer the effects of such costs on members' trade. Distance has long been a central variable explaining trade volumes, and typically has been interpreted as a measure of transport costs. However, the effect of distance probably measures more than transport costs, such as information costs. For instance, empirical work explaining bilateral foreign direct investment (FDI) flows also finds that distance has an economically significant effect on deterring such flows, even though theory suggests that measures of trade costs and FDI costs should have opposite effects on each others' flows (see Markusen 2002). Portes and Rey (2005) find empirically that distance also has a significant negative effect on portfolio flows, for which the transaction cost should be minimal.

Other sources of trade costs (which consume resources) include infrastructure, communication and foreign exchange costs. Limao and Venables (2001) find that infrastructure has a significant effect on trade volumes, with a decline in the level of infrastructure investment from the median level to the 75th percentile equivalent to a 2,166-mile (3,466-km) increase in sea distance travelled. Tang (2006), using various measures of information technology, finds that communication costs have a significant effect on the volume of trade. Economists have long thought that exchange rate variability and its associated uncertainty should impose a significant trade cost and deterrent to trade. However, a survey of studies reveals that the trade- volume effects are probably small (see Cote 1994).

With the use of the gravity equation in international trade, indirect estimates of the trade costs associated with national trade policies have proliferated. While a few empirical studies have looked at the effects of tariff rates explicitly on trade flows, the vast bulk have measured the presence or absence of (typically regional) economic integration agreements and currency unions on trade between country pairs. Many earlier studies using standard gravity equations found surprisingly small estimated effects on trade of arguably important trade agreements such as the Treaty of Rome (see Frankel 1997). However, more recent studies incorporating modern theoretical foundations for the gravity equation and econometric techniques suggest that such small estimates are probably due to a bias introduced by self-selection of countries into such agreements (see Baier and Bergstrand 2004, 2007). By contrast, work by Rose (2000) indicates that a currency union may have a very strong impact on trade between country pairs.

Estimates of trade costs have also been inferred indirectly using discrepancies in prices between countries. Engel and Rogers (1996) demonstrated that price variability of similar goods between US and Canadian cities is much greater than that between equidistant cities in the same country. Engel and Rogers (2001) showed that a 'real barriers' effect owing to incomplete market integration is present, but also that some dispersion could be explained by exchange rate variability and sticky price behaviour. Parsley and Wei (2001) showed that distance, unit-shipping costs, and exchange rate variability all contribute to dispersion of relative tradable goods prices across 96 cities in Japan and the United States. However, Crucini et al. (2005) used price data from European cities to show that goods markets – at least, those in Europe – may be much more integrated than earlier work showed.

See Also

- ▶ Currency Unions
- ▶ Factor Content of Trade
- ▶ Foreign Direct Investment
- ▶ Gravity Equation
- ▶ International Outsourcing
- ▶ Regional and Preferential Trade Agreements

Bibliography

Anderson, J.E. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69: 106–116.

Anderson, J.E., and E. van Wincoop. 2003. Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93: 170–192.

Anderson, J.E., and E. van Wincoop. 2004. Trade costs. *Journal of Economic Literature* 62: 691–751.

Baier, S.L., and J.H. Bergstrand. 2001. The growth of world trade. *Journal of International Economics* 53: 1–27.

Baier, S.L., and J.H. Bergstrand. 2004. Economic determinants of free trade agreements. *Journal of International Economics* 64: 29–64.

Baier, S.L., and J.H. Bergstrand. 2007. Do free trade agreements actually increase members' international trade? *Journal of International Economics* 71: 72–95.

Bergstrand, J.H. 1985. The gravity equation in international trade: Some microeconomic foundations and empirical evidence. *Review of Economics and Statistics* 67: 474–481.

Cote, A. 1994. Exchange rate volatility and trade: A survey. Working Paper, No. 94–5, Bank of Canada.

Crucini, M., C.I. Telmer, and M. Zachariadis. 2005. Understanding European real exchange rates. *American Economic Review* 95: 724–738.

Engel, C., and J.H. Rogers. 1996. How wide is the border? *American Economic Review* 86: 1112–1125.

Engel, C., and J.H. Rogers. 2001. Deviations from purchasing power parity: Causes and welfare costs. *Journal of International Economics* 55: 29–58.

Feenstra, R.C. 2004. *Advanced international trade: Theory and evidence*. Princeton: Princeton University Press.

Frankel, J.A. 1997. *Regional trading blocs*. Washington, DC: Institute for International Economics.

Hummels, D. 1999. Have international transportation costs declined? Unpublished manuscript, Purdue University.

Hummels, D. 2001. Toward a geography of trade costs. Unpublished manuscript, Purdue University.

Limao, N., and A.J. Venables. 2001. Infrastructure, geographical disadvantage, transport costs and trade. *World Bank Economic Review* 15: 451–479.

Markusen, J.R. 2002. *Multinational firms and the theory of international trade*. Cambridge, MA: MIT Press.

Obstfeld, M., and K. Rogoff. 2000. The six major puzzles in international macroeconomics: Is there a common cause? In *NBER macroeconomics annual*, ed. B.S. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.

Parsley, D.C., and S.-J. Wei. 2001. Explaining the border effect: The role of exchange rate variability, shipping costs, and geography. *Journal of International Economics* 55: 87–106.

Portes, R., and H. Rey. 2005. The determinants of cross-border equity flows. *Journal of International Economics* 65: 269–296.

Rose, A. 2000. One money, one market: Estimating the effect of common currencies on trade. *Economic Policy* 30: 7–45.

Tang, L. 2006. Communication costs and trade of differentiated goods. *Review of International Economics* 14: 54–68.

Trade Cycle

A. Medio

Keywords

Business cycle; Capital intensity; Chaos; Classical theory of the cycle; Entrepreneurship; Equilibrium; Equilibrium business cycle theory; Forced saving; Hawtrey, R. G.; Hayek, F. A. von; Hobson, J. A.; Innovation; Inventory investment; Juglar cycles; Kitchin cycles; Kitchin, J.; Kondratieff cycles; Long run and short run; Marginal efficiency of capital; Monetary theory of the cycle; Multiple equilibria; Multiplier–accelerator theory; Overinvestment theory of the cycle; Population growth; Stability of equilibrium; Technical change; Trade cycle; Underconsumption theory of the cycle; Uniqueness of equilibrium; Voluntary saving; Walras's Law; White noise

JEL Classifications

F1

The dynamics of capitalist economies are characterized by two facts: sustained growth of production and employment and wide oscillations of these magnitudes and the level of prices as well. This oscillatory behaviour of economic activity as

a whole is indeed the subject of trade cycle theory. The use of the word 'cycle', besides pointing to alternation of ups and downs, also suggests the idea that oscillations are somewhat regular.

Concerning the occurrence and regularity of the cycle there exist two entirely different positions. According to one line of thought, it is possible to explain it exogenously. The economic system by itself would not display any tendency to fluctuate regularly but for the influence of external cyclical impulses such as the alternation of seasons. A more sophisticated version of this principle recognizes that external impulses do not even have to be cyclical in order to induce regular fluctuations of the system. This approach, which is very old and recurrent among economists, reflects an essentially pessimistic view about their ability to explain a prominent feature of modern economies.

The opposite opinion holds that the generation and the persistence of cycles are totally or mainly endogenous to the economic system. This is an idea rather difficult to argue cogently. One can fairly easily represent a dynamic system characterized by a succession of expansions and contractions. It is a much harder task to define rigorously a system whose oscillations persist indefinitely, independent of external impulses, with amplitude and frequency determined solely by the structural parameters. Only recently has the economics profession acquired the necessary mathematical tools to solve this problem satisfactorily.

A possible third alternative – which possesses some elements of the other two – is to postulate that oscillations are normally damped and would therefore eventually die out, were they not continually revived by erratic shocks that resupply the system with the energy needed to sustain the cyclical motion. We shall return to this idea later since its discussion merges into that of some recent developments in economic dynamics.

The very existence of specific (and measurable) cycles of different lengths has been the object of heated discussions among economists. Joseph Schumpeter, the author of a monumental historical investigation on business cycles (1939), detected three main types of cycles. The shortest ones, named after the economist Joseph Kitchin, would last approximately three years. The intermediate ones (the Juglar cycles) would comprise three Kitchins and last approximately ten years. Finally, the 'long waves' (Kondratieff) would reflect major technological innovations and extend over 50–60 years.

Subsequent empirical research failed to find conclusive evidence of the actual existence of such regularities in the ups and downs of capitalist economies. And yet the recession of the 1970s has prompted a renewed interest in Schumpeter's work and one hears economists speak once again of the Kondratieff cycle.

Two considerations are here in point. First of all, the (courageous) formulations of hypotheses concerning the exact shape and duration of cycles and the subsequent discussions thereof provided fresh evidence as well as new and better tools of analysis, which greatly improved our knowledge of the subject, even when those hypotheses were eventually discarded. Secondly, since the 1930s economic agents' improved understanding of fluctuations, and the vastly increased public intervention to counteract their most negative consequences, have certainly modified both the economic mechanisms that produce those fluctuations and people's expectations about them. On this point too we shall have to return.

The discussion of long waves naturally leads one to consider trend in relation to cycle. Indeed the rising phase of a very long cycle may not be distinguishable over a certain period of time from sustained growth. The presence of a trend itself raises a number of problems. First of all, we would like to understand the economic forces that determine it. In most analyses of growth and cycle one assumes that trend is determined mainly by such long-run factors as population growth and technical progress. The latter, however, are usually represented by given functions of time, which is tantamount to admitting that we do not know much about them.

On the other hand, the causal relation may well run in the opposite direction: that is to say, in an economy characterized by sustained expansion, both population growth and productivity might be stimulated by general prosperity, the latter being determined by other factors. In most

economic systems there obviously exist 'hereditary factors' owing to which the 'long run' is resolved into a chain of 'short run' events. For instance, technical progress – at least that of the 'learning by doing' type – depends on the cumulative levels of production (or investment) over a more or less distant past. Moreover, both productivity and desired consumption display important 'ratchet effects', that is, they move more easily upwards than downwards. All these considerations suggest a dependence of trend on the cyclical path actually followed by the economy. However, a rigorous analysis of the process through which this influence is exerted is still lacking.

There also exists an inverse causal relation running from trend to cycle; that is, the characteristics of a cycle, in particular its amplitude and duration, are markedly different according whether the economy is in a phase of stagnation or prosperity.

Unfortunately, the question of the relation between cycle and trend remains to this day one of the many obscure points of economic theory. Most models tend to overcome this difficulty by assuming that short run (cycle) and long run (trend) may be dealt with separately and then re-combined in a 'cyclical growth model'. This procedure, however, requires rather formidable (and usually hidden) hypotheses; in particular, the relevant dynamic equations of the model must be linear, so that one may apply the principle of superposition of solutions of a dynamic model. The different authors' explanations of the trade cycle are better investigated in the wider context of their economic theories as a whole and nothing close to a comprehensive survey may be attempted here. We shall only try to put the contemporary discussions into a historical prospective, without which it would be difficult to understand what the various theories state and why.

Broadly speaking, we can identify three main phases in the development of cycle theory.

The first phase – the classical one – comprises the works of economists who wrote in the 18th century and in the first half of the 19th century. These writers did not provide a true scientific explanation of the cycle, but addressed certain basic questions whose understanding would prove essential to subsequent developments. In particular, classical economists debated the question of the stability of economic systems with a view to establishing whether capitalist economies possess an inherent tendency to equilibrium, that is, whether they are able to generate and maintain prices and quantities consistent with one another as well as with the structural parameters of the system. This problem, formulated by Adam Smith in terms of the 'invisible hand', was discussed in the first half of the 19th century mainly in relation to the possibility (or probability) of general crises of overproduction. In this context, Say (1803), Ricardo (1817) and Mill (1821) shared the opinion that production always creates its own demand and no general overproduction is therefore possible. Lauderdale (1804), Sismondi (1819) and especially Malthus (1820) dissented and pointed out that accumulation (saving) is not simply a redistribution of expenditure between consumption and investment goods. If incentive to invest is lacking or too weak, saving may well result in a general lack of expenditure and a consequent decline of production and employment.

The idea that capitalist economies have an intrinsic tendency to disequilibrium was also argued, in a very different context and with different overtones, by Karl Marx. Marx never produced a rigorous explanation of the cycle but his ideas provided inspiration to contemporary writers, in particular Kalecki and Goodwin. Marx, too, opposed the so-called Say's Law and pointed out that in a market economy, where purchases and sales are disjoint operations connected by the intermediation of money, discrepancies between demand and supply are always possible, not only in each individual sector but also in the economy as a whole. In volume 1 of *Das Kapital* (1867), Marx came close to formulating a self-consistent model of the cycle. Accumulation of capital, Marx argues, by reducing the rate of unemployment (the 'reserve industrial army'), pushes up wages (down profits), thus discouraging further investment. The ensuing recession brings about higher unemployment, leading to lower wages (higher profits) and therefore re-establishing the profitability of accumulation. This cyclical mechanism would be reformulated rigorously one century later by Richard Goodwin

(1967) and others, to provide a 'classical theory of the cycle' based on the interaction of capital accumulation and distribution of income. A second phase in the theory of the cycle, which we can locate between 1850 and 1930, may be termed 'modern' and among its forerunners we shall mention Juglar (1862), Jevons (1884) and Tugan-Baranowsky (1901). In this phase the cycle as such became an object of specific investigation. Economists then attempted to define hypotheses and construct theories to explain the 'true and fundamental' causes of economic fluctuations.

Broadly speaking, we can distinguish three main 'explanations' of the cycle, namely: (i) a monetary theory; (ii) an overinvestment theory, in its turn subdivided into a monetary and a real variation; (iii) an underconsumption theory. We shall briefly present the main ideas of the leading exponents of each group of theories. Separate consideration will be given to Keynes's contribution, which marks a watershed between modern and contemporary cycle theories.

The monetary theory of the trade cycle has been most forcefully argued by Ralph Hawtrey (1919). According to this author, the rising phase of the cycle is caused by credit expansion, realized mainly through a reduction of the rate of interest. This induces inventory accumulation by dealers, whose increased demand in turn stimulates producers' expenditure. The rise in demand for investment and consumption goods brings about a undesired *reduction* of the inventory–sales ratio to which dealers respond with further accumulation. A self-sustaining expansionary process ensues – possibly reinforced by secondary speculative waves – and will continue as long as monetary expansion goes on.

However, Hawtrey argues, insofar as the monetary system is constrained by a link between global liquidity and a real asset whose quantity is limited, monetary expansion must come to an end. Monetary flows, moreover, tend to generate fluctuations of real variables, owing to the lagged response of the demand for money to changes in income. Thus, in the ascending phase of the cycle, demand for money does not grow *pari passu* with expenditure and consumers' income. Consequently the banking system experiences net monetary inflows, which permit it to pursue an expansionary policy. However, as soon as demand for money catches up with income, banks' liquidity deteriorates, forcing them sooner or later to increase the rate of interest and squeeze credit. This initiates a downwards cumulative process. In a recession the lagged response of the demand for money will again act as a brake, slowing down the decline of the real output and employment. Eventually excess liquidity will reappear, the rate of interest will be reduced and the cycle will start anew.

The overinvestment theory of the cycle in its monetary version is perhaps best represented by Hayek's so-called 'concertina effect', as described in his book *Prices and Production* (1931).

Two ideas here play a crucial role. The first one is a typically neo-Austrian proposition stating that capital intensity (assumed to be unambiguously measurable) is an inverse function of the rate of interest. The second idea is based on the distinction between voluntary and forced saving. An increase in voluntary saving is accompanied by a reduction in the rate of interest and by an increase in capital intensity. While these adjustments take place, the system remains in equilibrium and no fluctuations arise. On the contrary, when saving is forced by an excessive credit expansion, investment is no longer constrained by *ex ante* saving. Owing to the unduly low rate of interest, the production process becomes too 'indirect' that is, there will be an excessive development of those stages of production which are more removed from the final, consumption stage.

In sum, according to Hayek's theory, the structure of demand and in particular the distribution between investment and consumption goods is determined by the propensity to save, whereas the structure of production is a function of the rate of interest. In equilibrium the rate of interest is fixed so as to make those two structures consistent. Excess credit causes an overproduction of capital goods, which must eventually manifest itself through increased profitability in the consumption good sector and corresponding losses in the investment sector. The latter will therefore experience a crisis that will turn the boom into a recession.

It is interesting to observe that in both Hawtrey's and Hayek's theories the culprit for the recession is the banking system, which keeps the rate of interest too low during expansion. However, for Hayek this leads to an undue 'lengthening' of the production process, bound to be reversed when the supply of money ceases to grow at an excessive rate. For Hawtrey, a low rate of interest provokes an excessive rise in global demand vis-à-vis the available stock of money.

A real version of the overinvestment theory of the cycle was put forward by Wicksell (1907) and Schumpeter, whose ideas on this subject are best summarized in the already quoted treatise (1939). These authors shared the view that growth and cycle are intrinsically related, and the main theoretical problem in this context is to explain why economic expansion does not take place smoothly, 'as trees grow', but why it occurs in leaps and bounds. Wicksell and Schumpeter also agreed that the oscillatory behaviour of capitalist economies is related to the process of innovation through which the employment of limited quantities of primary factors (labour and land) results in increasingly greater amounts of consumption and investment goods. Essentially the trade cycle depends on the fact that innovations, that is, the introduction of new techniques, new products and new markets, are not distributed uniformly in time, but take place in a discontinuous manner, in groups or 'swarms'.

Schumpeter especially emphasized the role of entrepreneurial activity in the generation of business cycles. Since innovation implies a break of routine, it cannot occur smoothly but requires a minimum critical amount of energy to overcome inertia. Such energy is provided by exceptional individuals who possess the courage, strength and imagination necessary to 'do things differently'. Once these few pioneers have opened the way, many others will follow to share in the extra-defeating process, made available by innovation. This is a self-defeating process, though. Insofar as the new methods products have been absorbed by the market, prices and profits will fall, terminating the boom and reversing the direction of the cycle. Secondary waves many amplify the oscillations, bringing about overoptimistic booms followed by severe slumps.

For Schumpeter and Wicksell alike, however, not everything about recession is bad, provided the worst consequences of deep depressions can be avoided. Recession is indeed a phase of adjustment during which innovations are 'digested' and leads the economy to an intrinsically superior stage. Recession, so to speak, realizes what the boom had promised: a permanently increased flow of commodities, reduced costs, entrepreneurial profits transformed into higher incomes of the other social classes. Schumpeter insisted upon the Darwinian function of (normal) recessions, during which 'lame ducks' are eliminated and only the stronger and more efficient survive. Similar arguments would be employed in the 1980s on both sides of the Atlantic Ocean to justify severe anti-inflationary policies.

The best-known exponent of underconsumption theories of the cycle was undoubtedly Hobson (1922). His theory of investment did not differ much from that of Wicksell and Schumpeter and he shared their view that investment opportunities are basically determined by the needs of development, which are in turn governed by technical changes and population growth.

According to Hobson, though, depressions are caused by insufficient expenditure, which in capitalist economies arises from a skew distribution of income. Increases in income are accompanied by more than proportional increases in saving, leading to overinvestment first and overproduction later. Hobson did not believe in the efficacy of the traditional remedies to overproduction, namely a reduction in the rate of interest and in the level of prices. As concerns the former, he was of the opinion that saving responded little, if at all, to changes in the interest rate. On the other hand, changes in prices are too sluggish to counteract undesired changes in real variables. Instead, actual economies eliminate excess saving in the most inefficient and painful way, that is, through depression and unemployment.

Hobson's arguments are clear anticipation of certain Keynesian ideas which would become very popular a few decades later.

It may be noticed that both the overinvestment and the underconsumption theories locate the origin of the crises in a 'vertical' imbalance between the investment and consumption sectors, an idea which had already been discussed by Marx, Rosa Luxemburg and Tugan-Baranowsky. Those theories differ from one another concerning which of the two sectors is overexpanded, the investment sector according to the former, the consumption sector to the latter. Both of them, moreover, could be comprised under the more general label of overcapitalization theories, but the one (underconsumption) argues that there are too many capital goods (and consequently too many consumption goods) vis-à-vis global demand, the other (overinvestment) maintains that there is excess accumulation vis-à-vis saving. It follows that policy recommendations suggested by the supporters of these theories conflict: a reduction of consumption according to overinvestment; a redistribution of income leading to greater consumption according to underconsumption theory.

Keynes's own theory of the cycle – as distinguished from Keynesian theories – constitutes a *trait d'union* between the modern and the contemporary phase. Even if all the elements necessary to build a true model of the cycle exist in Keynes's work, this task was left to his followers. We shall describe here the essence of Keynes's argument as it appears in the *General Theory* (1936).

At an abstract level, Keynes was fully aware that cyclical motion must result from alternations in the relative strength of expansive and contractive forces, to wit that the cycle must be nonlinear. The changes in the balance between expansive and contractive forces may take place smoothly or abruptly, the latter case being more frequently in a boom, the former in a slump. These ideas would be further developed by younger economists inspired by Keynes, in particular by Kaldor and Goodwin.

For Keynes the trade cycle is a complex mechanism but its most important manifestations are the fluctuations of the marginal efficiency of capital, which is determined by psychological as well as economic considerations, and primarily depends on abundance or scarcity of capital goods, their cost and expectations about their future returns. Most often a recession is precipitated by a turnaround of expectations, which, to Keynes, is a much more decisive factor than the increase in the rate of interest sometimes associated with it.

In the last part of the boom, entrepreneurs' optimism offsets all the other unfavourable circumstances: excess capital goods, increasing costs and a high rate of interest. Estimates of future returns to assets are distorted and exaggerated by ignorance, speculation, and vested interests of financial intermediaries and asset-owners. When finally the overoptimistic forecasts are falsified by facts, there follow excessive and even catastrophic readjustments. The increase in liquidity preference that usually accompanies the decrease in the marginal efficiency of capital aggravates the situation and nullifies the effects of expansive monetary measures.

At this point, Keynes contrasts his own theory with that based on overinvestment and observes that the latter is an ambiguous term. If overinvestment means that capital goods in general are so abundant that no investment project can be found whose expected return could justify the cost, then, according to Keynes, this is a rare occurrence even at the peak of a boom. Instead – Keynes continues – investment is excessive either in the sense that actual returns are lower than those on the expectations of which the decision to invest has been taken, or in the sense that severe unemployment makes investment superfluous. It follows that in a recession the right remedy is to reduce, not to increase, the rate of interest.

On the other hand, Keynes thought that the essential propositions of the underconsumption theory were substantially correct. Under capitalist rules of the game, the volume of investment is not controlled, let alone planned, and depends on the vagaries of the marginal efficiency of capital, with a rate of interest systematically kept above a conventional minimum. In this situation, in order to maintain high rates of employment it may well be necessary to stimulate consumption. Keynes's only criticism of the underconsumption theory is its neglect of the possibility of stimulating investment directly. But this – according to Keynes – is a matter of expediency rather than theory.

The modern phase of the cycle theory was characterized by a rich theoretical investigation which led to a deeper understanding of the dynamics of capitalist economies. It did not result, however, in the production of true models, that is, sets of rigorously formulated propositions containing the (necessary and) sufficient conditions to represent in idealized form the cyclical behaviour of the economy. The 'contemporary' phase of cycle theory – from the 1930s onwards – does not consist so much in the search for new explanations of the 'enigma of the cycle' (Wicksell 1907) as in the attempt to analyse in a rigorous manner concepts and ideas put forward by earlier writers in an intuitive form.

The foundations of the mathematical theory of the trade cycle were laid down mostly in the 1930s, and the early works in this field (Frisch 1933; Tinbergen 1959: Kalecki 1935, 1943; and Samuelson 1939) were characterized by a distinctly Keynesian flavour. Keynes's own writings – and of course even more so the world crisis – had convinced a growing number of economists that the free functioning of the market was more likely to lead to an oscillatory behaviour of income, employment and prices, than to full employment equilibrium.

Multiplier–accelerator theory, in different versions and with different refinements, rapidly became predominant in the decades immediately before and after the Second World War.

In the 1960s, however, the economics profession's attention turned away from the problem of the cycle. Several reasons contributed to this change, but four of them seem to stand out as crucial.

First of all, economic events after the Second World War seemed to suggest that business cycles had become obsolete. After a phase of settlement following the turmoil of the war, the world economy (or at least that of the major industrialized countries) seemed to have entered an epoch of sustained growth without (or with only minor) fluctuations. Economists, and especially the younger ones, were quick to respond to the current social and political mood.

Secondly, the prevailing models of the cycle, those of Keynesian inspiration, suffered from a major drawback. The solution of such models consists of fluctuations, perfectly regular in the sense that a certain periodic orbit, once established, repeats itself over and over again.

In such a situation, economic agents would sooner or later notice the periodic character of the dynamics of the system and learn to calculate the amplitude and frequency of the cycles. This in turn would lead to a revision of their expectations. The behavioural hypotheses of the model – on which the cyclical motion of the system depends – would no longer be tenable and the model itself would have to be reformulated. Incidentally, this criticism of the deterministic models of the cycle is perhaps the most important element of truth in the theory of rational expectations.

Thirdly, historical data on the main economic variables do not confirm the periodic regularity predicted by the model.

Finally, and partly in response to the theoretical difficulties mentioned above, there has recently been a revival of what we would call the 'static prejudice' in economics, whose most explicit expression is perhaps 'equilibrium business cycle theory'. In a nutshell, this theory describes the working of the economy by means of a model whose deterministic part is characterized by a unique, stable equilibrium. A stochastic part is added to it which depends on imperfect information of economic agents, whose decisions are therefore mistaken. With rational expectations these errors are 'white noise', that is, they are normally distributed above and below the optimal (equilibrium) values. The resulting fluctuations of the system are non-periodic but bounded and their amplitude and frequency can be estimated statistically. This approach is reminiscent of certain early ideas first developed in the 1920s and 1930s by Slutsky (1937), Hotelling (1927), Yule (1927), Frisch (1933) and later by Kalecki (1954), to explain the persistence of the cycle. However – as Hicks commented long ago (1950) – when the random factors 'explain' a substantial part of the deviations of a system from its equilibrium position, the proposed theory amounts to a confession of ignorance.

Mentioning a 'static prejudice' in economics raises a few methodological questions, discussion

of which is perhaps the best way to introduce the recent developments in the theory of the cycle (and, more generally in economic dynamics) and to conclude this essay.

In spite of the great practical importance of the problems discussed and the high intellectual quality of the results obtained, cycle theory has generally been regarded as an interesting but as a whole marginal branch of economics, whose exponents as such rarely managed to hold centre stage in professional debates.

The great theoretical systems developed in the 18th and 19th centuries, for example, those of Ricardo and of Walras, were more suited to the explanation of interdependence of economic variables, rather than their evolution in time. This problem is better formulated by means of a system of algebraic equations, whose solution (not necessarily unique) is a set of values of the variables – typically prices of commodities and quantities produced – which is consistent with the postulated relationships and the exogenously fixed parameters. The equations themselves define certain equilibrium conditions, for example, equality of demand and supply, or uniformity of the rate of profits. Uniqueness and stability of equilibrium are deemed to be desirable properties of the model, in a descriptive as well as in normative sense. Indeed only stable systems are observable. Multiple equilibria, on the other hand, are not satisfactory for two reasons. First of all, in the general case they are alternatively stable and unstable; secondly, the multiplicity of equilibria introduces a certain amount of relativism as far as their optimality properties are concerned.

Discussion of stability must bring (and historically has brought) dynamic considerations into the picture. Early writers provided intuitive, but not very cogent stories arguing that the system 'tended to' or 'gravitated towards the natural or equilibrium values and prices'. Walras took a step forward, providing a principle (Walras's Law) which shows that, under the postulated maximizing behaviour, disequilibria of notional demands and supplies cannot be entirely haphazardous, but most take place according to certain rules (that is, the value of total excess demand must be equal to zero). This led to rather optimistic considerations as far as stability of competitive equilibrium was concerned, but was a far cry from a rigorous statement of the problem.

A deeper understanding of the intricacies of non-equilibrium behaviour of dynamical systems and more powerful mathematical techniques were required. Fundamental progress along this line was made in the interwar period (Hicks 1939) and more so during and after the Second World War, when certain necessary mathematical results became available to economists. (The list of important contributions is too long to be quoted exhaustively here: we shall only mention Samuelson 1941–1942; Metzler 1945; Morishima 1952; Arrow and Hurwicz 1958.)

The conclusions were rather dismaying though, since a rigorous analysis showed that sufficient stability conditions entailed the most heroic assumptions on the specifications of the system.

Economists' overwhelming preoccupation with equilibrium explains their neglect of cycle theory. Indeed cycle theory must by definition concern itself with off-equilibrium states of the system. Empirical observation suggests that the economy is not normally in equilibrium, but fluctuates without ever coming to rest or, if we except relatively rare historical occurrences, without 'exploding'. Is this restlessness only the result of random external shocks, or is it a structural characteristic of the system, owing to the operation of endogenous mechanisms? In dealing with this problem the cycle theorist is naturally led to posing typically dynamic questions such as 'what are economic agents' reactions to non-equilibrium, and therefore unsatisfactory situations?'; or, more generally, 'what laws of motion govern the system, starting from a given, generally off-equilibrium state?'

An exact formulation of this problem required an even more sophisticated analytical apparatus than was the case with static theory. Its natural mathematical setup is a system of differential (or difference) equations, whose solution is a set of functions of time which, given the initial conditions, describes, so to speak, the history of the variables under consideration. The fact that the

relevant part of theory of differential equations became available to economists at a rather late date is perhaps a further explanation of their preference for static rather than dynamic problems.

In discussing the problem of fluctuations, economists must soon run up against the question of linearity. Most systems, as described by economists' models, are linear; that is, the coefficients that appear in those models do not depend on the variables or their derivatives with respect to time. The crucial importance of this assumption can be appreciated by considering how it affects the analysis of stability. From the point of view of linear analysis, equilibrium is either stable or unstable. A disturbance is therefore followed by a return to rest in the former case, or by an indefinite increase in the magnitude of the deviation in the latter. But this is a very rough description of the behaviour of an economic system. Linear analysis, by its very nature, completely disregards the effect of the size of disturbances. Indeed an economy may be stable with respect to small deviations from equilibrium, but not so when those deviations are large. On the other hand, an economy may be locally unstable, in the sense than it does not show any tendency to return to its equilibrium position when subjected to small shocks, but it may possess self-correcting mechanisms which only operate far from equilibrium. In either case, the conclusions reached by means of a linearization around the equilibrium point would be misleading.

The linear approximation is unsatisfactory not only because it provides a simplified and therefore distorted picture of reality. A much more fundamental shortcoming is that there exist certain phenomena (economic or otherwise) which cannot be idealized at all by means of linear models. In particular, it is well known that a system of linear differential (or difference) equations, if structurally stable, cannot describe sustained oscillations, that is, oscillations that do not expire or explode.

The discovery that certain basic questions of economic dynamics, in particular persistent cycles, could not be tackled effectively by means of linear models led an increasing number of economists to make use of nonlinear methods of analysis. In so doing, they experienced what Poincaré – to whom modern dynamic analysis owes more than to anybody else – had described several decades earlier. The study of the cycle had proved precious to the analyst, all the more so since it constituted a first necessary step into the mysterious and hitherto inaccessible realm of non-equilibrium dynamics.

At that point the phrase 'cycle theory' became an elliptical expression designating not a particular problem, but economic dynamics as a whole or, more appropriately, a method for investigating dynamic processes in economics. But from the analysis of fluctuations economists have obtained much more than a number of new methods and mathematical tools. The most aware of them have undergone a true cultural revolution that Frisch already in the early 1930s (1933) thought would be as important to economics as the transition from classical to quantum mechanics had been to physics.

Equilibrium states, stable and unstable and even limit cycles have now been revealed as rather special configurations in a much more complex and morphologically rich theoretical universe. As soon as the linearity assumption has been dropped, even a simple model may exhibit a very complicated behaviour.

The objection may be raised that this is very nice, but what does it have to do with real systems, in our case real economies? On the contrary, recent developments in dynamic theory have finally provided economists with tools of analysis such that theoretical results are, at least qualitatively, comparable with empirical observations. Take, for example, so-called chaotic behaviour. This characterizes a large class of dynamic models, economic or otherwise and, in common parlance, describes irregular fluctuations of a type that has so far been exclusively associated with random (and therefore essentially unexplained) disturbances. After all, what could be more realistic than irregular but bounded fluctuations of income, employment and prices, and what more unrealistic than a stable unique equilibrium point?

Clearly cycle theory and, more generally, economic dynamics is in a state of transition. To produce theoretically meaningful and socially relevant models of real systems, economists must fulfil two main requirements:

(i) to define mechanisms of adjustment that realistically describe economic agents' behaviour in disequilibrium;
(ii) to employ techniques of analysis suitable to study the dynamic systems resulting from the operation of those mechanisms.

Recent developments have contributed substantially to solve some of the problems in (i). Instead, off-equilibrium behaviour of economic agents remains to this day a rather fuzzy area of economic investigation.

See Also

► Aggregate Demand Theory

Bibliography

Arrow, K.J., and L. Hurwicz. 1958. On the stability of the competitive equilibrium. *Econometrica* 26: 522–552.

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honour of Gustav Cassel*. London: G. Allen & Unwin.

Goodwin, R.M. 1950. A non-linear theory of the cycle. *Review of Economics and Statistics* 32: 316–320.

Goodwin, R.M. 1951. The non-linear accelerator and the persistence of business cycles. *Econometrica* 19: 1–17.

Goodwin, R.M. 1953. Econometrics in business cycle analysis. In *Readings in business cycles and national income*, ed. R.V. Clemence and A.H. Hansen. New York: Norton.

Goodwin, R.M. 1955. A model of cyclical growth. In *The business cycle in the post-war world*, ed. E. Lundberg. London: Macmillan.

Goodwin, R.M. 1967. A growth cycle. In *Socialism, capitalism and economic growth*, ed. C.H. Feinstein. Cambridge: Cambridge University Press.

Hawtrey, R.G. 1919. *Currency and credit*. London: Longmans.

Hayek, F. 1931. *Prices and production*. London: G. Routledge & Sons.

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Hicks, J.R. 1950. *A contribution to the theory of the trade cycle*. Oxford: Clarendon Press.

Hobson, J.A. 1922. *The economics of unemployment*. London: G. Allen & Unwin.

Hotelling, H. 1927. Differential equations subject to error, and population estimates. *Journal of the American Statistical Association* 22: 283–314.

Jevons, W.S. 1884. *Investigations in currency and finance*. London: Macmillan.

Juglar, C. 1862. *Des crises commerciales, et leur retour périodique en France, en Angleterre, et aux Etats-Unis*. Paris: Guillaumin.

Kaldor, N. 1940. A model of the trade cycle. *Economic Journal* 50: 78–92.

Kalecki, M. 1935. A macrodynamic theory of business cycles. *Econometrica* 3: 327–344.

Kalecki, M. 1943. *Studies in economic dynamics*. London: G. Allen & Unwin.

Kalecki, M. 1954. *Theory of economic dynamics*. London: G. Allen & Unwin.

Kalecki, M. 1968. Trend and business cycle reconsidered. *Economic Journal* 78 (June): 263–76; corrigendum, September, 729.

Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economy 1933–1970*. Cambridge: Cambridge University Press.

Lauderdale, J.M. 1804. *An inquiry into the nature and origin of public wealth and into the means and causes of its increase.* Edinburgh: A. Constable & Co./London: Hurst, Robinson & Co.

Lucas, R. 1981. *Studies in business cycle theory.* Oxford: Basil Blackwell.

Malthus, T.R. 1820. *Principles of political economy.* London: John Murray.

Marx, K.H. 1867. *Capital*, vol. 1. New York: Penguin Books, 1976.

Metzler, L.A. 1945. The stability of multiple markets: The Hicks conditions. *Econometrica* 13: 277–292.

Mill, J. 1821. *Elements of political economy.* London: Baldwin, Cradock & Joy.

Morishima, M. 1952. On the laws of change of price-system in an economy which contains complementary commodities. *Osaka Economic Papers* 1 (May): 101–113.

Ricardo, D. 1817. *On the principles of political economy and taxation*. London: John Murray.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics* 21 (May): 75–78.

Samuelson, P.A. 1941–1942. The stability of equilibrium. Pt. I: Comparative statics and dynamics; Pt. II: Linear and non-linear systems. *Econometrica* 9 (April 1941): 97–120; 10 (January 1942): 1–25.

Say, J.B. 1803. *Traité d'économie politique*. Paris: Déterville.

Schumpeter, J.A. 1939. *Business cycles: A theoretical, historical and statistical analysis of the capitalist process*. New York/London: McGraw-Hill.

Sismondi, J.C.L. 1819. *Nouveaux principes d'économie politique*. Paris: Delaunay.

Slutsky, E. 1937. The summation of random causes as the source of cyclical processes. *Econometrica* 5 (April): 105–146.

Tinbergen, J. 1959. *Selected papers*, ed. L.H. Klaassen. Amsterdam: North-Holland.

Tugan-Baranowsky, M. 1901. *Studien zur Theorie und Geschichte der Handelskrisen in England.* Jena.

Wicksell, K. 1907. Krisornas gata. Statskonomisk Tidsskrift 21, 255–84. Trans. C.G. Uhr as 'The Enigma of Business Cycles', *International Economic Papers* no. 3 (1953), 58–74.

Yule, G.U. 1927. On a method of investigating periodicities in disturbed series. *Philosophical Transactions of the Royal Society of London* Series A 226 (April): 267–98.

Trade Policy, Political Economy of

Devashish Mitra

Abstract

This area of research tries, through the introduction of politics in economic models, to explain the existence and the extent of anti-trade bias in trade policy. The two main approaches, namely, the median-voter approach and the special-interest approach are surveyed. Certain applications of these approaches to policy issues, such as trade agreements, the issue of reciprocity versus unilateralism in trade policy, regionalism versus multilateralism, hysteresis in trade policy and the choice of policy instruments, are discussed. Finally, the empirical literature on the political economy of trade policy is surveyed. The new literature that employs a more 'structural' approach is emphasized.

Keywords

Congestion problem; Cournot oligopoly; Customs unions; Deadweight loss; Free trade; Free trade areas; Free-rider problem; Heckscher–Ohlin trade theory; Hysteresis in trade policy; Intermediate goods; International trade; Lobby formation; Lobbying; Majority rule; Median voter model; Monopolistic competition; Multilateralism in trade policy; Non-tariff barriers; Optimal obfuscation principle; Political competition; Political economy; Political support function; Political-contributions model; Progressive and regressive taxation; Proportional representation; Protection; Reciprocity in trade policy; Regional and preferential trade agreements; Risk sharing; Single-peaked preferences; Special-interest politics; Specific-factors trade theory; Tariff-formation function; Tariffs; Tariffs vs. subsidies; Terms of trade; Trade agreements; Trade policy, political economy of; Trade-diverting vs. trade-creating bilateral agreements; Unilateralism in trade policy; World Trade Organization

JEL Classifications

F1

While economists clearly understand the benefits of free trade, they have always found it difficult to explain departures from it in the real world. Most of these departures are in the direction of limiting the volume of trade. In trying to explain the existence and the extent of this anti-trade bias in trade policy, trade economists have introduced politics in their economic models. Parts of this political-economy literature have also tried to explain why policy instruments that are more efficient than trade policy and can achieve the same political and economic objectives are not often used. An important empirical contribution of the political-economy literature has been to uncover the main determinants of cross-country and cross-industry variations in protection.

Modelling Approaches

Median-Voter Approach

Political economy models of trade are of two main types. One of them adopts the majority voting approach. Such models are called 'median voter' models in the literature. Preferences are assumed to be 'single peaked' and conditions are imposed such that the most preferred policy of each individual is monotonic in a certain characteristic. Then, with other individual characteristics held constant across the population, the tariff chosen under two-candidate electoral competition is the median voter's most preferred tariff. The median voter here is the median individual in the economy when ranked according to the

characteristic under consideration. Mayer (1984) applies this median-voter principle to the Heckscher–Ohlin and specific- factors trade models. In the Heckscher–Ohlin case, the political economy equilibrium tariff is the most-preferred tariff of the median individual in the economy-wide ranking of the ratio of capital to labour ownership. If this median individual's capital to labour ratio is less than the economy's overall capital to labour ratio – that is, if the asset distribution in the economy is unequal – the equilibrium trade policy is different from free trade and is one that redistributes income from capital to labour – pro-trade in a labour-abundant economy and anti-trade in a capital-abundant economy.

Special-Interest Politics

The other type of political economy model in the trade literature focuses on 'specialinterest' politics. The first papers to model lobbying explicitly in the trade arena were by Findlay and Wellisz (1982) and Feenstra and Bhagwati (1982). The Findlay–Wellisz model is a two-sector model in which production in each sector is carried out using a factor of production specific to that sector – land for food production and capital for manufactures – and an economy-wide general factor, namely, labour. Both types of specific factor owners are fully organized politically and they lobby against each other. This simple model shows the existence of an equilibrium tariff determined through the Nash interaction between the two opposing groups. The government is modelled very indirectly through a tariff formation function which is increasing in the amount of labour devoted to lobbying by the import-competing specific factor and decreasing in labour used in lobbying by the specific-factor owners in the export sector.

While only labour is used as an input in lobbying in the Findlay–Wellisz model, both capital and labour are used as inputs into lobbying in the Feenstra–Bhagwati model. However, only one sector is assumed to be politically active in the model. Unlike in the Findlay–Wellisz model, the government in the Feenstra–Bhagwati model is not a monolithic entity but has a two-layered structure. While one layer is a clearing house for lobbies, the other cares about social welfare. The tariff is determined through an interaction between the two layers.

Another approach to modelling 'special-interest' trade politics is the 'political support function' approach pioneered by Hillman (1989). (Some of the classification terminology here is borrowed from Rodrik 1995, to which the interested reader is referred for a detailed typology of political-economy models. See also Helpman 2002, for an analytical survey within a unified framework.) Under such an approach, the government's objective function, also called the political support function, incorporates its preferential treatment of each organized industry as well as the cost of protecting this industry given by the excess burden on society. Van Long and Vousden (1991) use a specific form of Hillman's political support function which is linear in the welfare levels of different types of specific-factor owners, with different weights being assigned to different factors.

Magee et al. (1989) explicitly model electoral competition. They use a two-sector, two-factor Heckscher–Ohlin set-up with two political parties – one pro-trade and another pro-protection – and two lobbies – one representing capital and the other labour. Lobbies contribute to their respective favoured political parties to maximize their chances of winning elections. Policy platforms here are chosen prior to decisions on campaign contributions.

The special-interest approach has evolved from the simple Findlay–Wellisz 'tariff-formation function' approach to the state-of-the-art Grossman and Helpman (1994) 'political-contributions' model. The latter is path-breaking for several reasons. First, it is multi-sectoral. Second, it provides micro-foundations for the behaviour of organized lobbies and politicians. A 'menu-auctions' approach is used in modelling policy bidding by interest groups. Multiple principals, namely, the various organized lobbies, try to influence the common agent, namely the government. The government's objective function is a weighted sum of political contributions and aggregate welfare, while each lobby maximizes its welfare net of political contributions.

Most importantly, especially from the empirical angle, the level of protection for each industry is derived as an estimable function of industry characteristics and other political and economic factors. Protection to organized sectors is negatively related to import penetration and the (absolute value of) import demand elasticity, while protection to unorganized sectors is positively related to these two variables. With everything else held constant, organized sectors are granted higher protection than unorganized sectors.

While Grossman and Helpman in their models take the existence of organized lobbies as given, Mitra (1999) extends their framework to endogenize lobby formation. He shows that we are closest to free trade when the government cares too little or too much about aggregate welfare relative to political contributions. While the former leads to the formation of a large number of mutually opposing lobbies, the latter situation is one where hardly any lobbies get formed. Mitra also shows that a higher concentration of asset ownership in the economy leads to the formation of a larger number of organized lobbies representing sectors that are heavily protected. Magee (2002) analyses a single lobby's organization problem in the context of the collection of political contributions in a repeated game setting. (See also Pecorino 1998, for an analysis of the same issue with a tariff-formation function approach.)

Theoretical Applications

Trade Agreements

The first important theoretical application we discuss is the issue of trade agreements. Using their political-contributions approach, Grossman and Helpman (1995a) analyse trade policy in a setting with two large countries, where they show an additional terms-of-trade component in the tariff expression in a non-cooperative setting. This component gets eliminated in a cooperative setting of international trade negotiations, and the relative size of protection in any sector in the two countries then depends on the relative political power of the same industry in the two countries. Thus there is a rationale for 'trade talks' as opposed to 'trade wars'. Using what is very close to a 'political-support' function approach, Bagwell and Staiger (1996, 1999) show that, even when political economy considerations are taken into account, the only rationale for (reciprocal) trade agreements is the elimination of terms-of-trade externalities. They use this approach to develop a rationale for the General Agreement on Tariffs and Trade (GATT)/World Trade Organization (WTO) and its different rules. (See Bagwell and Staiger 2002, for an in-depth discussion.)

The next natural question then is whether free trade agreements are of any value to countries whose actions have no impact on the international terms of trade. Maggi and Rodriguez-Clare (1998) have a political economy explanation for the unilateral commitment to free trade agreements by small countries. Their setting is one in which owners of capital first decide in which sector to invest, and then those who invest in a particular sector (the import-competing sector) lobby the government for protection. The lobbying is modelled as a Nash bargaining game between the lobby and the government. While the lobby at least compensates the government for the deadweight losses generated in the second stage, it may not compensate the government for the welfare loss through the inter-sectoral misallocation of capital in the first stage in the expectation of protection in the second stage. In such a situation, it is possible that a government will commit to a free trade agreement in a prior stage 'zero'.

Mitra (2002) builds on the Maggi–Rodriguez-Clare version of the Grossman–Helpman framework, augmenting it with the decision to incur fixed costs (to build relationships with politicians in power and/or to form a lobby) prior to the actual lobbying, but, importantly, not providing room for any capital mobility. However, the main result of the Maggi–Rodriguez-Clare model goes through even in this newly modified set-up. This is the result that generally governments with low bargaining power with respect to domestic lobbies are the ones that precommit to free trade agreements.

Grossman and Helpman (1995b) have provided a detailed analysis of politicaleconomy factors responsible for the emergence of free trade agreements. Using their 'political-contributions' approach, they show that such agreements between two countries are impossible if in every sector one country has a higher tariff than the other. These agreements might be politically feasible only when tariffs on some goods are higher in one country while other tariffs are higher in its partner country. The possibility of exclusion of certain sectors from the trade agreement also raises the chances that the agreement will be signed.

Reciprocity and Unilateralism in Trade Policy

I now move to the issue of reciprocity and unilateralism in trade policy. Bagwell and Staiger (1996, 1999, 2002) have analysed the issue of reciprocal trade liberalization in considerable detail in both bilateral and multilateral settings (see also Hillman and Moser 1996). In their models, reciprocity is a way of eliminating terms-of-trade externalities in the setting of trade policy. While considerable work has been done on the role of reciprocity in trade policy, the causal interaction between unilateral and reciprocal trade liberalization has been a somewhat neglected issue. Krishna and Mitra (2005) modify the Mitra (1999) lobby-formation framework to study exactly this link. (See Bhagwati 1990, for an early informal discussion of this idea. See also Coates and Ludema 2001, for an alternative channel based on risk sharing through which unilateralism induces reciprocity.) While reciprocal reduction in trade barriers can reasonably be expected to occur in contexts involving trade negotiations between countries, Krishna and Mitra examine instead the question of whether unilateral trade liberalization by one country can induce reciprocal liberalization by its partner *in the absence of any communication or negotiations* between the two countries. In this context, they show that unilateral liberalization by one country can affect the political economy equilibrium in the partner country through the formation of an export lobby there, in a manner that induces it to liberalize trade.

The Political Economy of Regionalism Versus Multilateralism

An important question raised by Bhagwati (1993, 1994) in several of his writings is whether regionalism is a 'stumbling block' or a 'stepping stone' to multilateralism. (For a purely economic answer that relies on coordination failure based on sector-specific sunk costs and 'friction' in trade negotiation, see McLaren 2002.) Levy (1997) uses a Heckscher–Ohlin set-up with monopolistic competition and a median- voter approach to address this issue. He finds that bilateral agreements between countries similar in factor endowments result in the subsequent blocking of multilateral trade agreements. He also finds that bilateral agreements can never increase the political support for multilateralism. Krishna (1998) addresses the same issue in a political economy set-up where profits get a much greater weight than other components of welfare in the government's objective function (political-support function approach). The set-up is one of Cournot oligopoly. He finds greater political support for trade-diverting bilateral agreements (regionalism) than for trade-creating ones. Such agreements can also make previously feasible multilateral agreements politically infeasible. This effect turns out to be increasing in the magnitude of the trade diversion that takes place under bilateralism.

Free Trade Areas Versus Customs Unions

Next I move to the determinants of the actual shape or form a preferential trading arrangement will take. Panagariya and Findlay (1996) study the choice between a customs union and a free-trade area in the context of how they affect lobbying activity and the structure of external tariffs. Using a tariff-formation function approach, they focus on the free-rider problem in lobbying in the case of customs union arising from the requirement of a common external tariff. Richardson (1994) is similar in spirit and finds the same free-rider effect under a customs union, due to which free-trade areas are preferred by import-competing producers.

McLaren (2004) takes a different approach to the choice between a free trade area and a customs union. He analyses what he calls the 'dynamics of

political influence'. As the external tariff is common across members of a customs union, it has to be set jointly by all the members, which can be done only if an agreement is reached among them. This makes the external tariff relatively less reversible under a customs union than under a free-trade area. It is this relative irreversibility that McLaren focuses on, even though he abstracts from the uniformity aspect. Using a political contributions approach with Nash bargaining between the capitalists and the government, he finds that a customs union is more likely when the government has a short lifespan and firms have a long lifespan. This is because the more permanent nature of trade policy under a customs union requires the upfront payment of contributions, which is the government's share in the present value of the stream of surpluses generated over time.

The Choice of Policy Instruments

The next important application concerns the choice of policy instruments. One of the major propositions of the theory of commercial policy is that, if distortions or policy goals are not directly trade-related, then direct subsidies are more efficient than tariffs (Bhagwati and Ramaswami 1963; Johnson1965; Bhagwati 1971). One simple explanation for the existence of tariffs despite their low efficiency is that they generate revenues while subsidies use them up (see for instance Bhagwati and Ramaswami 1963).

However, in a Bhagwati and Srinivasan (1980, 1982) framework of fully competitive revenue seeking, subsidies can be preferable to tariffs as they are not subject to revenue seeking.

Rodrik (1986) is the first author to look at this issue by endogenizing policy. He does this by using a simplified version of the Findlay–Wellisz model. He argues that, since tariffs are general to an industry and subsidies can, in principle, be firm- specific, the free-rider problem in lobbying for tariffs may result in a smaller level of endogenous tariffs than endogenous subsidies, thereby possibly reversing the conventional welfare ranking of tariffs and subsidies. Mitra (2000) argues that, even from the point of view of the -import-competing firms, tariffs may be preferable to subsidies since lobbying in the latter may face a congestion problem while in the former the free-rider problem may offset the congestion problem.

This issue of the choice of instruments is also addressed by Grossman and Helpman (1994). They argue that, when the policy instrument used for redistribution is more efficient, it creates greater competition among lobbies and thus results in a larger proportion of the surplus in the hands of the government. Therefore, lobbies themselves may want to tie the hands of the government to using relatively inefficient instruments. Wilson (1990) makes a similar argument in a model with electoral competition where he shows that a higher efficiency of redistributive instruments leads to more contributions and more transfers in equilibrium.

The choice between tariffs and subsidies has also been considered in a voting framework in a series of papers by Mayer and Riezman (1987, 1989, 1990). They show that tariffs can be chosen in equilibrium outcome when voters differ along dimensions other than factor endowments, such as tastes and preferences, treatment under income taxes, and so forth. Besides, income tax progressivity might mean that the cost of financing subsidies is borne unevenly, which might lead some individuals to prefer tariffs whose costs are more evenly distributed.

Feenstra and Lewis (1991) argue that tariffs are informationally more efficient than subsidies. When the world price of importables falls, a tariff equivalent to this decline will compensate the losers without making others worse off relative to the initial situation. This will be possible without any knowledge on the part of the government of individual production and consumption levels. Magee, Brock and Young (1989) propose another information-based explanation, which they call the principle of 'optimal obfuscation'. Indirect policies such as tariffs are less observable by those who bear its costs.

Staiger and Tabellini (1987) argue that, when governments provide surprise protection to those hurt by world price fluctuations, the time-inconsistency problem might be less severe with more inefficient policies.

Hysteresis in Trade Policy

A model that helps us understand status quo bias in trade policy is Fernandez and Rodrik (1991). They consider a two-sector economy that initially has a certain given tariff on its imports. Eliminating this tariff will result in a movement of workers from the import-competing sector to the export sector. What is *ex ante* unknown is which of the workers initially in the import-competing sector will be successful in moving to the export sector. All the workers who are in the export sector right from the beginning will gain, while those who are always in the import-competing sector and remain there after the reforms will lose. Another group that gains is the group of movers from the contracting import-competing sector to the expanding export sector. Suppose 30 per cent of the population is in the export sector and 70 per cent in the import-competing sector to start with. After the reforms, let us suppose that this split is 60 per cent and 40 per cent respectively. This means that 60 per cent of the population will gain *ex post* from the reform. While 30 per cent who are initially in the export sector know for sure *ex ante* they are going to benefit, the remaining 70 per cent do not know which 30 per cent out of them will lose and which 40 per cent will gain. If they know for sure that the loss incurred by the losing 40 per cent is greater than the gain to the remaining 30 per cent, then all the voters who are initially in the import-competing sector will vote against the reform. Due to the individual-specific uncertainty faced by workers in the import-competing sector, each of them will vote on the basis of an expected loss, arising from the fact that losers in this sector lose much more than gainers in that sector gain. Thus, even though *ex post* a majority gain from the reforms, *ex ante* a majority of the workers vote against the trade reforms. However, if a dictator or an international financial institution forces a reform upon these people, it will not be reversed since as we know in this case *ex post* there is going to be majority support for the reforms.

Beyond the Monolithic Government

In the existing political economy literature on trade policy determination, a single, monolithic policymaker is often assumed. Until very recently, the paper by Feenstra and Bhagwati (1982) was the only exception. McLaren and Karabay (2004) make a departure from such a simple structure to study trade policy setting in the presence of parliamentary or congressional institutions. They also incorporate electoral competition between political parties and show that in their setting the equilibrium tariff is the optimum of the median voter in the median district. They find that the relationship between the likelihood of import protection and the geographical concentration of import-competing interests is non-monotonic, with a maximum occurring at moderate levels of concentration. Too much concentration leads to a control of too few seats, while too much dispersion leads to no control of any seats.

A paper by Grossman and Helpman (2005) allows actual policy formation to be the interplay between the policy platform announced by the party leadership and the actions of individual legislators who want to maximize political success. Maximization of political success involves resolving the trade-off between conforming to the party platform and making one's constituents happy, thereby resulting in a deviation of 'policy reality' from 'policy rhetoric'. The authors find a protectionist bias when the legislature operates under majority rule. This bias is increasing in the geographical concentration of assets and capital market imperfections, and is decreasing in party discipline.

Empirical Evidence

The Old Empirical Literature

The empirical literature in this area has evolved from being highly 'reduced form' and atheoretical to being fairly 'structural' and guided by tight theoretical models. Important papers in the earlier literature include Caves (1976), Saunders (1980), Ray (1981), Marvel and Ray (1983), Ray (1991) and Trefler (1993). (See Rodrik 1995, for a detailed survey of this literature.) The main finding of this early empirical literature is that protection is higher for sectors that are labour-intensive, low-skill and low-wage, for consumer-goods industries, for industries facing high import

penetration, where geographical concentration of production is high but that of consumers is low, and in sectors with low levels of intra-industry trade. (For an examination of the cross-national variation in average protection levels across industrialized countries, see Mansfield and Busch 1995. They find that non-tariff barriers are increasing in country size, unemployment rate and number of parliamentary constituencies, and are higher for countries that use proportional representation as their electoral system.)

The New Empirical Literature

In the Heckscher–Ohlin version of the Mayer median-voter model, a simple comparative static exercise produces the result that a rise in asset inequality will make trade policy more pro-trade in a labour-abundant economy and more protectionist in a capital-abundant economy. Dutt and Mitra (2002) find strong support for this result using cross-country data on inequality, capital-abundance and diverse measures of protection. (In this context, it is also important to mention Milner and Kubota 2005, who use a median-voter approach to empirically investigate the relationship between democratization and trade reforms in developing countries.)

Dutt and Mitra (2005) also perform a cross-country empirical investigation of the role of political ideology in trade policy determination. They use a political-support function approach within a two-sector, two-factor Heckscher–Ohlin model (see Milner and Judkins 2004, on this issue. Hiscox (2001) studies six Western nations to look at how historically the nature and structure of partisanship on trade issues change over time and depend on the extent of inter-sectoral factor mobility. Hiscox (2002) looks at the same question exclusively for the United States, analysing major pieces of congressional trade legislation between 1824 and 1994.

Two empirical papers, Goldberg and Maggi (1999) and Gawande and Bandyopadhyay (2000), estimate the Grossman–Helpman 'Protection for Sale' tariff expressions using industry-level data from the United States. The two papers are similar in the questions they address, but are somewhat different in the details of their approaches. While Goldberg and Maggi restrict their focus to the protection expressions, Gawande and Bandyopadhyay concentrate more on the lobbying aspects and the determinants of the magnitude of contributions. Goldberg and Maggi use the basic Grossman–Helpman framework, while Gawande and Bandyopadhyay introduce intermediate goods. The econometric specifications are therefore somewhat different in the two papers. However, the results in the two papers are very similar. Both confirm empirically the Grossman–Helpman prediction regarding the relationship of protection to import penetration and import-demand elasticity. With everything else held constant, organized sectors are granted higher protection than unorganized sectors. Both papers find that the weight on aggregate welfare in the government's objective function is several times higher than that on contributions. Also, the estimates of the proportion of the population organized are very high in both papers.

Mitra et al. (2002) and McCalman (2004) obtain similarly high parameter estimates of the Grossman–Helpman model for Turkey and Australia respectively. An interesting result that comes out of the empirical exercise by Mitra, Thomakos and Ulubasoglu is that the relative weight on aggregate welfare was higher in the democratic regimes than under the dictatorial regimes in Turkey.

Gawande et al. (2006) empirically investigate 'the susceptibility of government policies to lobbying by foreigners'. Using a new data-set on foreign political activity in the United States, they investigate the empirical relationship between trade protection and lobbying activity. Their theoretical framework is an extension of the 'Protection for Sale' model to include foreign lobbies. They find that foreign lobbying activity has significantly affected US trade barriers in a negative direction, as predicted by their model. They conclude: 'If the policy outcome absent any lobbying by foreigners is characterized by welfare- reducing trade barriers, lobbying by foreigners may result in reductions in such barriers and raise consumer surplus (and possibly improve welfare).'

In another empirical application through an extension of the Grossman–Helpman model, Gawande and Krishna (2005) investigate the effects

of lobbying competition between upstream and downstream producers for US trade policy. Their parameter estimates are a significant improvement over those in the earlier literature even though they do not completely resolve the puzzle.

Thus, we see that the political economy literature on trade policy has evolved a great deal on both the theoretical and the empirical sides, as well as in terms of the complexity of applications it can handle.

See Also

- ▶ 'Political Economy'
- ▶ Policy Reform, Political Economy of
- ▶ Political Competition
- ▶ Political Institutions, Economic Approaches to

Bibliography

Bagwell, K., and R. Staiger. 1996. *Reciprocal trade liberalization*, Working paper, vol. 5488. Cambridge, MA: NBER.

Bagwell, K., and R. Staiger. 1999. An economic theory of GATT. *American Economic Review* 89: 215–248.

Bagwell, K., and R. Staiger. 2002. *The economics of the world trading system*. Cambridge, MA/London: MIT Press.

Bhagwati, J. 1971. The generalized theory of distortions and welfare. In *Trade, balance of payments and growth*, ed. J. Bhagwati et al. Amsterdam: North-Holland.

Bhagwati, J. 1990. Aggressive unilateralism. In *Aggressive unilateralism*, ed. J. Bhagwati and H. Patrick. Ann Arbor: University of Michigan Press.

Bhagwati, J. 1993. Regionalism and multilateralism: An overview. In *New dimensions in regional integration*, ed. A. Panagariya and J. De Melo. Washington, DC: World Bank.

Bhagwati, J. 1994. Threats to the world trading system: Income distribution and the selfish hegemon. *Journal of International Affairs* 48: 279–285.

Bhagwati, J., and V. Ramaswami. 1963. Domestic distortions, tariffs and the theory of the optimum subsidy. *Journal of Political Economy* 71: 44–50.

Bhagwati, J., and T. Srinivasan. 1980. Revenue seeking: A generalization of the theory of tariffs. *Journal of Political Economy* 88: 1069–1087.

Bhagwati, J., and T. Srinivasan. 1982. The welfare consequences of directly unproductive profit-seeking (DUP) lobbying activities: Price versus quantity distortions. *Journal of International Economics* 13: 33–44.

Caves, R. 1976. Economic models of political choice: Canada's tariff structure. *Canadian Journal of Economics* 9: 278–300.

Coates, D., and R. Ludema. 2001. A theory of trade policy leadership. *Journal of Development Economics* 65: 1–29.

Dutt, P., and D. Mitra. 2002. Endogenous trade policy through majority voting: An empirical investigation. *Journal of International Economics* 58: 107–133.

Dutt, P., and D. Mitra. 2005. Political ideology and endogenous trade policy: An empirical investigation. *Review of Economics and Statistics* 87: 59–72.

Feenstra, R., and J. Bhagwati. 1982. Tariff seeking and the efficient tariff. In *Import competition and response*, ed. J. Bhagwati. Chicago/London: University of Chicago Press.

Feenstra, R., and T. Lewis. 1991. Distributing the gains from trade with incomplete information. *Economics and Politics* 3: 29–40.

Fernandez, R., and D. Rodrik. 1991. Resistance to reform: Status-quo bias in the presence of individual-specific uncertainty. *American Economic Review* 81: 1146–1154.

Findlay, R., and S. Wellisz. 1982. Endogenous tariffs, the political economy of trade restrictions and welfare. In *Import Competition and Response*, ed. J. Bhagwati. Chicago/London: University of Chicago Press.

Gawande, K., and S. Bandyopadhyay. 2000. Is protection for sale? A test of the Grossman–Helpman theory of endogenous protection. *Review of Economics and Statistics* 82: 139–152.

Gawande, K., and P. Krishna. 2003. The political economy of trade policy: Empirical approaches. In *Handbook of international trade*, ed. J. Harrigan and E. Kwan Choi. Malden, MA: Basil Blackwell.

Gawande, K., and P. Krishna. 2005. *Lobbying competition over US trade policy*, Working paper, vol. 11371. Cambridge, MA: NBER.

Gawande, K., P. Krishna, and M. Robbins. 2006. Foreign lobbies and US trade policy. *Review of Economics and Statistics* 88(3): 563–571.

Goldberg, P., and G. Maggi. 1999. Protection for sale: An empirical investigation. *American Economic Review* 89: 1135–1155.

Grossman, G., and E. Helpman. 1994. Protection for sale. *American Economic Review* 84: 833–850.

Grossman, G., and E. Helpman. 1995a. Trade wars and trade talks. *Journal of Political Economy* 103: 675–708.

Grossman, G., and E. Helpman. 1995b. The politics of free trade agreements. *American Economic Review* 85: 667–690.

Grossman, G., and E. Helpman. 2005. A protectionist bias in majoritarian politics. *Quarterly Journal of Economics* 120: 1239–1282.

Helpman, E. 2002. Politics and trade policy. In *Interest groups and trade policy*, ed. G. Grossman and E. Helpman. Princeton: Princeton University Press.

Hillman, A. 1989. *The political economy of protection*. Chur: Harwood Academic Publishers.

Hillman, A., and P. Moser. 1996. Trade liberalization as politically optimal exchange of market access. In *The new transatlantic economy*, ed. M. Canzoneri, W. Ethier, and V. Grilli. Cambridge: Cambridge University Press.

Hiscox, M. 2001. *International trade and political conflict: Commerce, coalitions and mobility.* Princeton: Princeton University Press.

Hiscox, M. 2002. Commerce, coalitions, and factor mobility: Evidence from congressional votes on trade legislation. *American Political Science Review* 96: 593–608.

Johnson, H. 1965. Optimal trade interventions in the presence of domestic distortions. In *Trade growth and balance of payments*, ed. R. Caves and H. Johnson. Amsterdam: North-Holland.

Krishna, P. 1998. Regionalism and multilateralism: A political economy approach. *Quarterly Journal of Economics* 113: 227–251.

Krishna, P., and D. Mitra. 2005. Reciprocated unilateralism in trade policy. *Journal of International Economics* 65: 461–487.

Levy, P. 1997. A political-economic analysis of free-trade agreements. *American Economic Review* 87: 506–519.

Magee, C. 2002. Endogenous trade policy and lobby formation: An application to the free-rider problem. *Journal of International Economics* 57: 449–471.

Magee, S., W. Brock, and L. Young. 1989. *Black hole tariffs and endogenous policy theory.* Cambridge/New York: Cambridge University Press.

Maggi, G., and A. Rodriguez-Clare. 1998. The value of trade agreements in the presence of political pressures. *Journal of Political Economy* 106: 574–601.

Mansfield, E., and M. Busch. 1995. The political economy of trade barriers: A crossnational analysis. *International Organization* 49: 723–749.

Marvel, H., and E. Ray. 1983. The Kennedy Round: Evidence on the regulation of trade in the US. *American Economic Review* 73: 190–197.

Mayer, W. 1984. Endogenous tariff formation. *American Economic Review* 74: 970–985.

Mayer, W., and R. Riezman. 1987. Endogenous choice of trade policy instruments. *Journal of International Economics* 23: 377–381.

Mayer, W., and R. Riezman. 1989. Tariff formation in a multidimensional voting model. *Economics and Politics* 1: 61–79.

Mayer, W., and R. Riezman. 1990. Voter preferences for trade policy instruments. *Economics and Politics* 2: 259–273.

McCalman, P. 2004. Protection for sale and trade liberalization: An empirical investigation. *Review of International Economics* 12: 81–94.

McLaren, J. 2002. A theory of insidious regionalism. *Quarterly Journal of Economics* 117: 571–608.

McLaren, J. 2004. *Free trade agreements, customs unions and the dynamics of political influence*. Mimeo: University of Virginia.

McLaren, J., and B. Karabay. 2004. Trade policy making by an assembly. In *Political economy of trade, aid and foreign investment policies*, ed. D. Mitra and A. Panagariya. Amsterdam: Elsevier.

Milner, H., and B. Judkins. 2004. Partisanship, Trade policy, and globalization: Is there a left–right divide on trade policy? *International Studies Quarterly* 48: 95–119.

Milner, H., and K. Kubota. 2005. Why the move to free trade? Democracy and trade policy in the developing countries. *International Organization* 59: 107–143.

Mitra, D. 1999. Endogenous lobby formation and endogenous protection: A long run model of trade policy determination. *American Economic Review* 89: 1116–1134.

Mitra, D. 2000. On the endogenous choice between protection and promotion. *Economics and Politics* 12: 33–52.

Mitra, D. 2002. Endogenous political organization and the value of trade agreements. *Journal of International Economics* 57: 473–485.

Mitra, D., D. Thomakos, and M. Ulubasoglu. 2002. 'Protection for sale' in a developing country: Democracy vs. dictatorship. *Review of Economics and Statistics* 84: 497–508.

Panagariya, A., and R. Findlay. 1996. A political-economy analysis of free trade areas and customs union. In *The political economy of trade reform: Essays in Honor of Jagdish Bhagwati*, ed. R. Feenstra, D. Irvin, and G. Grossman. Cambridge, MA: MIT Press.

Pecorino, P. 1998. Is there a free-rider problem in lobbying? Endogenous tariffs, trigger strategies and the number of firms. *American Economic Review* 88: 652–660.

Ray, E. 1981. The determinants of tariff and non-tariff restriction in the United States. *Journal of Political Economy* 89: 105–121.

Ray, E. 1991. Protection of manufactures in the United States. In *Global protectionism: Is the US playing on a level field?* ed. D. Greenaway. London: Macmillan.

Richardson, M. 1994. Why a free trade area? The tariff also rises. *Economics and Politics* 6: 79–96.

Rodrik, D. 1986. Tariffs, subsidies and welfare with endogenous policy. *Journal of International Economics* 21: 285–296.

Rodrik, D. 1995. Political economy of trade policy. In *Handbook of international economics*, vol. 3, ed. G. Grossman and K. Rogoff. Amsterdam: North-Holland.

Saunders, R. 1980. The political economy of effective protection in Canada's manufacturing sector. *Canadian Journal of Economics* 13: 340–348.

Staiger, R., and G. Tabellini. 1987. Discretionary trade policy and excessive protection. *American Economic Review* 77: 823–837.

Trefler, D. 1993. Trade liberalization and the theory of endogenous protection. *Journal of Political Economy* 101: 138–160.

Van Long, N., and N. Vousden. 1991. Protectionist responses and declining industries. *Journal of International Economics* 30: 87–103.
Wilson, J. 1990. Are efficiency improvements in government transfer policies self defeating in equilibrium? *Economics and Politics* 2: 241–258.

Trade Subsidies

Murray C. Kemp

That it may be in the interest of a prince or nation to subsidize foreign trade is an ancient doctrine. However, the manner in which subsidization has been justified and the means by which it has been effected have changed radically over the years. In the 17th and 18th centuries, the subsidization of exports was a corollary of the general Mercantilist doctrines of the time (Viner 1937); and the British Navigation Acts were defended, even by Adam Smith (1776), as ensuring that England would be adequately provided with ships and sailors in time of war. In the 19th century, Alexander Hamilton and the economists List, J.S. Mill and Bastable argued that industries which in the face of foreign competition are unprofitable but which are capable of learning might qualify for temporary support (see Kemp 1974, for a modern statement of this 'infant industry' doctrine). And some 20th-century economists have advocated export subsidies as a means of alleviating unemployment.

Subsidization may be direct or indirect. Direct subsidies to trade are simply negative taxes and have repercussions which are, in all details, the opposite of those associated with taxes. An indirect subsidy, on the other hand, may take the form of a 'tax holiday', easy credit, cheap power or free infrastructure. The amount of the subsidy may bear a simple relationship of constant proportionality to the volume of trade (a *specific* subsidy) or to the value of trade (an *ad valorem* subsidy). Or it may be determined by a more complicated formula, as with the variable levy on the agricultural imports of the European Economic Community, the purpose of which is to stabilize internal producers' prices; for a comparison of the welfare implications of the variable levy and alternative stabilization devices, see Young and Kemp (1982).

Formal economic analysis has focused on direct trade subsidies which, as we have noticed, are simply negative taxes. Much of that analysis has dealt with highly simplified worlds with just two primary factors and two traded goods. The following propositions have emerged.

(a) To each *ad valorem* rate of export subsidy there corresponds an *ad valorem* rate of import subsidy with the same impact on world and domestic price ratios (Lerner's (1936) Symmetry Theorem).

(b) A subsidy to imports or exports will normally turn the terms of international trade against the policy-making country; but exceptions are possible, as noted by Marshall (1926), Lerner (1936) and Kemp (1966).

(c) A subsidy to imports or exports will normally turn the domestic price ratio against the subsidized good and therefore will normally turn the distribution of income against whichever factor is used relatively intensively in the subsidized industry; but Lerner (1936) has noted that exceptions are possible.

However, in recent years attention has moved to a more policy-relevant range of questions. Working with models which accommodate any number of traded and non-traded goods, economists have considered the implications for national welfare of prescribed patterns of change in taxes and subsidies on trade. Of particular interest is the discovery that, under certain conditions, a small country will benefit from a uniform proportionate reduction in all specific taxes and subsidies and from a reduction in that tax or subsidy which has the largest absolute value to the level of the next largest tax or subsidy (see Lloyd 1974; Hatta 1977; Fukushima 1979, 1981).

Economists have been much occupied with the implications of trade subsidies for national welfare. However, it should not be inferred from their choice of research topic that prevailing trade subsidies are, on the whole, designed by legislatures to promote the national interest. Most subsidies are monuments to past and present sectional interests.

See Also

► Effective Protection

Bibliography

Fukushima, T. 1979. Tariff structure, nontraded goods and the theory of piecemeal policy recommendations. *International Economic Review* 20(2): 427–435.

Fukushima, T. 1981. A dynamic quantity adjustment process in a small open economy, and welfare effects of tariff changes. *Journal of International Economics* 11(4): 513–529.

Hatta, T. 1977. A recommendation for a better tariff structure. *Econometrica* 45(8): 1859–1869.

Kemp, M.C. 1966. Note on a Marshallian conjecture. *Quarterly Journal of Economics* 80: 481–484.

Kemp, M.C. 1974. Learning by doing: formal tests for intervention in an open economy. *Keio Economic Studies* 11(1): 1–7.

Lerner, A.P. 1936. The symmetry between import and export taxes. *Economica* 3: 306–313.

Lloyd, P.J. 1974. A more general theory of price distortions in open economies. *Journal of International Economics* 4(4): 365–386.

Marshall, A. 1926. Memorandum on fiscal policy. In *Official papers by Alfred Marshall*, ed. J.M. Keynes. London: Macmillan.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Modern Library. Reprinted, 1937.

Viner, J. 1937. *Studies in the theory of international trade*. London: George Allen & Unwin.

Young, L., and M.C. Kemp. 1982. On the optimal stabilization of internal producers' prices in international trade. *International Economic Review* 23(1): 123–141.

Trade Unions

Henry Phelps Brown

When they formulated their classic definition of a trade union, Sidney and Beatrice Webb had in view the long struggle of groups of English workers to maintain associations that could stand up to employers and gain acceptance by the community. 'A trade union,' they said, 'is a continuous association of wage earners for the purpose of maintaining or improving the conditions of their working lives' (1894, p. 1). An economist starting from the assumption of the ultimate rationality of decisions is likely to see the trade union as a cartel or monopoly intended to maximize the benefits of its members. An intermediate view recognizes that men and women join trade unions for reasons that arise out of imperfections of the labour market. Because of the slow response of employment to lower labour cost, the job seekers in any one district will be confronted at a given time with a limited number of jobs: if then they exceed that number, even by one, and compete with each other by underbidding, the wage can be brought down to a limit set by bare subsistence or the level of support in unemployment.

But even suppose that the numbers of vacancies and applicants match exactly: then if the employer and an individual applicant cannot reach agreement on the rate for the job, so that for the time being the employer lacks a workman and the applicant has no job or pay, which is in the greater trouble? As Adam Smith said, 'In the long run the workman may be as necessary to his master as his master is to him; but the necessity is not so immediate.' The applicants here are evidently unable to move away readily to other employers: those with whom they are dealing are monopsonists or oligopsonists. Against this, they try to maintain a monopoly. They agree to hold out for a minimum in common. They want to keep up the price of their work by limiting the supply. They also want to safeguard their jobs against a drop in demand – they aim to establish a property in jobs. To these ends they defend lines of demarcation, within which they have the sole right to work, or they allow only approved entrants, in limited numbers, to acquire certain skills, or to be recruited for certain purposes. The defensive object of preventing their rates being undercut or their labour being displaced by outsiders merges here into the calculated purpose of pushing up their earnings by restricting supply.

In modern Western economies, trade union membership has also been maintained or extended, especially among white collar workers; by the need to renegotiate the pay of all employees to compensate for changes in the cost of living; the addition of an improvement factor in real

terms has also been regarded as defensive from the point of view of any one group, which would otherwise fall behind the others. Beyond these issues of the rate of pay and job security are those that arise at the place of work. People join trade unions to secure protection against discrimination and arbitrary treatment by management, and the negotiation and observance of a code governing discipline, grievance procedure, promotion, redundancy, the pace of work, and the like.

Factors Affecting Membership

These forces making for trade union membership have arisen and taken effect only in certain conditions. Where trade unions emerged, their form and function differed widely in different societies. In the Western democracies, the proportion of employees unionized has varied widely over time and between countries (Bain and Price 1960); sometimes it has fallen even against the trend of economic growth, notably in the USA since the 1960s. In full employment, and in places where the individual had access to a number of alternative employers, or to natural resources like the American open frontier, he would feel able to fend for himself. The absence of observed falls in wage rates down the centuries (Phelps Brown and Hopkins 1981) implies that custom and tacit understandings can maintain rates in the absence of overt trade unionism. Individuals whose qualifications, temperaments, and entries into employment interest them in personal advancement are not likely to become trade unionists; but these factors deterring clerical, administrative and managerial employees from membership have been offset by the growth of offices in size and impersonality, and the need of staff to negotiate frequent salary rises to offset inflation. The ability of manual workers to form their own trade unions has depended upon leaders coming forward from their ranks who were literate, upright, and skilful in administration; the workers themselves must be able to keep up a subscription, and have the discipline to sustain a stoppage. Where those conditions are lacking, as in much of the Third World, trade unions tend to be organized by outsiders, often a political party.

In all countries, the ability of trade unions to maintain themselves and function depends on the provisions of the law and their application in the courts: landmarks here were the immunity from civil liability conferred on British trade unions in 1906, and the promotion of trade unionism and collective bargaining by American legislation in the 1930s. Linked with this is the attitude of the employers: whereas those in France, Germany and the United States generally felt themselves justified, down to 1914 and sometimes later, in resisting trade unionism, many British employers had come to accept it as a means of stabilizing industrial relations. In the Soviet-type economies, discontent with the conditions of employment leading to combined action can result only in a political revolt: trade unions exist by name, but only to administer social benefits and maintain the control of the party within the establishment.

Trade unions thus have to be viewed in their local variety and historical setting. 'Where we expected to find an economic thread for a treatise,' the Webbs wrote in the Preface to their *History of Trade Unionism* (1894), 'we found a spider's web; and from that moment we recognized that what we had first to write was not a treatise, but a history.'

Trade Unions As Monopolies

None the less, there are economic threads to be followed through. One is the effect of the trade union on the relative pay of its members. Here the theory of the monopoly power of the trade union directs attention to the elasticity of substitution between the members' labour and other factors of production, and to the elasticity of demand for the product. Substantially, much depends on the possibility of the labour being replaced by equipment, and of the trade union gaining control of this if it were introduced. It is in the firms and industries themselves most strongly placed in the market, and able to retain ample margins, that trade unions are likely to maintain levels of pay above those obtaining elsewhere for similar grades of

labour. The employers concerned are thus paying what seems more than the supply price of labour to them, and the differences found in surveys in the rates paid even in adjacent firms suggest that this is so; the trade unions may be said to share in the monopoly power of the employers. They may also acquire monopoly power directly by restricting supply and by forcing demand.

Craft unions have restricted supply by limiting the numbers of apprentices; when a trade is being organized for the first time, attacks on non-members who are continuing to work for less than the union rate serve either to exclude or recruit them; and this shades into the general purpose of the rule that no one shall undertake work of a certain kind unless he or she holds a union card, which serves more for recruitment than for exclusion. The pre-entry closed shop provides the most complete control. A trade union that organizes all the existing workers in an industry has to reckon with the possibility of the market being invaded by the products of non-members newly employed elsewhere – except in those cases where of its nature the produce must be supplied in the place where it is consumed.

Trade unions force demand by rules preventing work being taken away from their members, such as compositors' work being done by advertisers, or builders' work by the makers of pre-fabricated components; by stopping other workers doing jobs in the territory to which they claim exclusive rights; and by resisting the application of labour-saving equipment to their own work. Many restrictive prices are intended to maintain or increase the input of labour per unit of output.

The monopoly power of a group of workers who form an essential link in a chain of production but account for a small part of the whole cost, appears great. Adam Smith instanced the half dozen woolcombers who were needed to keep a thousand spinners and weavers at work. Marshall asked why the bricklayers of his day did not get 'an enormous rise' by pushing their own rate up. This power is in fact limited by the employer's powers of resistance. He may redesign process or product, so as to by-pass the labour in question; he may put the work out to subcontract, or import components; at the limit, he may move the whole operation to a location where the trade union is not in control. His resistance to a claim by the union will also be stiffened by his knowledge that other groups in his employ will have regard to relativities, and will base their own claims on concessions he makes to the union.

Collective Bargaining

The most widely available use of monopoly power is the pushing up of the rate of pay by bargaining, which leaves it to employers to restrict the supply by the limitation of the number they engage at the higher rate. If we consider in the first place a negotiation whose effects are largely confined to the immediate parties, bargaining power proper may be defined as the power to inflict loss by withholding consent. It is understandable that if two parties cannot agree upon the terms of an agreement to work together, they should suspend operations meanwhile. But this suspension is not a merely negative act, for it puts each party into difficulties. Workers are left without pay. Craft unions have often had funds from which to issue strike pay; other unions, needing to keep subscriptions low, pay none, but have sometimes been able to maintain long strikes none the less with the aid of contributions from other unionists and the public. There has been a risk of the vacant places being filled by disloyal members of the union, or by imported blacklegs who will be kept on when the dispute is settled. These difficulties increase the longer the stoppage goes on. But so do those of the employer. There is the immediate loss of profitable operation, and in some industries this cannot in the nature of things be made good by increased output when work is resumed. There is the likelihood that customers will resort to other suppliers meanwhile, and the possibility that some of them will never return. Firms that have been unprofitable, though on that account they cannot easily afford a rise, may not however be able to hold out against settling for one, because of their attenuated cash flow. The actual experience of increasing difficulty makes the parties willing to modify the terms for which they stood out when the stoppage began: there is convergence, and

they reach agreement. Such at least is a natural interpretation of the observation that most stoppages have ended in a compromise. Reflecting on this, J.R. Hicks inferred (1932, chapter 7) that if the parties estimated each other's powers of resistance accurately beforehand, there would be no stoppage, but agreement would be reached at once on the terms reached only at the end of a stoppage that occurs when the parties do not know those powers, or misconceive them, and find out the facts by painful experience.

That most agreements are reached without a stoppage does not mean that bargaining power is not exerted. But more enters into the reaching of an agreement than bargaining power. The matter to be negotiated is the terms and conditions on which a joint activity is to be carried on by the parties in future, and this is not, like the price of a horse, to be haggled over between two people who may never deal with each other again: the relation between parties who continue to be indispensable to one another is more nearly matrimonial. The parties are therefore open to influence by the thought of what is fair and reasonable in the terms on which they can work together. Trade unionists may be moved by the aim, not of receiving the greatest possible gain, but of obtaining what is justly due to them, or of righting a wrong. Where justice is at stake they will fight without weighing the cost against the gain.

Another consideration in their conduct of a negotiation is their determination to avoid subservience. They refuse to accept the force of the remark that the improvements in terms achieved by a strike will not make good the wages lost in the strike until after many years: for that is an argument to show that the employer's superior resources should always oblige the workers to accept his terms. Bargaining may turn again into warfare, in which trade unionists whose blood is up will make sacrifices according to no maximizing calculus, and will attack blacklegs with patriots' hatred for a traitor.

So far, the bargaining power of the trade union has been considered as if it were exerted by one of two parties facing each other in isolation; but the power of many unions is enhanced by the impact of their strikes on third parties and on the community. The third parties who are most likely to be disturbed by the stoppage of the employer's activity, and interested in his reaching an early settlement, are the firms who supply him with substantial parts of their own output, and those who depend on him for supplies that they cannot readily replace from stock or from other sources. A trade union that can withhold the supply of an essential product or service from a whole region can force the intervention of the government.

In 1893 the power of the English Miners' Federation to cut off much of the country's heat, light and inland transport brought about what was unthinkable a short time before – the intervention of Government to effect a settlement. When the French railwaymen went on strike the Government broke the strike by mobilizing them for military service. A strike of the British miners that threatened to bring the whole country to a halt in 1912 was settled by an Act of Parliament that gave the miners much of what they had claimed. Where the Government has to settle a national emergency dispute, it cannot force the trade unionists to resume work on terms that they reject as unfair and unreasonable, but it can apply a substantial coercive force to the employers.

Control of essential supplies and services offered certain trade unions great power in this way, and the Triple Alliance of miners, transport workers and railwaymen was formed in Great Britain to exploit it; but the Government for its part built up a detailed organization, held in reserve against an emergency, for the maintenance of supplies. In the USA, the Taft–Hartley Act of 1947 provided that in a strike that creates a national emergency the President might take the business concerned into public possession for eighty days, during which the employees must return to work while a fact-finding board reported on the circumstances of the dispute. With the extension of trade unionism in the public sector and in services in Great Britain, the object of strikes has shifted from inflicting loss on the employer to demonstrating discontent by disrupting the activities of the community, and inflicting hardship on the parents of schoolchildren or on invalids or commuters.

Trade Unions and the Law

The bargaining power of trade unions depends upon legal privilege. Employers may refuse to recognize a trade union unless the law obliges them to do so. In a strike, labour is commonly withdrawn in breach of the individual worker's contract of employment; losses are usually inflicted on third parties. In the USA, employers were able to inhibit many forms of trade union action by obtaining injunctions against them from the courts, until the Norris-La Guardia Act of 1932. If those who suffer damages are able to bring civil actions to recover them, most strikes will be impossible: British trade unions have operated under the shelter of immunities that were given outright statutory form by the Trade Disputes Act 1906. Many strikes, again, will not be effective unless pickets are posted to turn back men and women who want to go on working, or stop supplies moving: the effectiveness of legal provisions designed to regulate picketing depends on the possibility and practice of enforcement. Not only the activities but the very existence of a trade union, as a combination in restraint of trade, are anomalous in a country whose common law protects the freedom of the individual to use his labour and property. In these countries the law has found a place for the trade union by way of large exception, rather than by the conferment of delimited rights.

The close bearing of the law on trade union activities has led the trade unions to bring pressure to bear on the legislature. The entering of representations on particular measures was the original purpose of the British Trades Union Congress, and the policy of the American Federation of Labour under Gompers. In later years the British trade unions have become the principal financial support of the Labour Party, and the American leadership has become associated with the Democratic Party. A main reason for association between European trade unions and a political party is the sharing of social principles and ideals, and in France and Italy different groups of trade unions are linked with different parties.

The Bargaining Area

Bargaining power cannot be considered apart from the bargaining area within which it is exerted. What that area shall be in a given case is the outcome of historical factors. Sometimes the initiative in shaping the present area has been taken by employers, sometimes by trade unionists. American employers, perhaps because they were highly individualistic and competitive, have generally been loath to associate, even for the legitimate purpose of collective bargaining, and the plant contract has predominated. The British tradition has been that of the craft union that has tried to make one rate obtain for all engagements, and maintain it through times of slack trade; here the wider the front that could he held, the better, and trade union policy drew together the major employers of each district. Through World War I this extended to industry-wide bargaining. 'Putting a floor under competition' throughout an industry in that way was a step towards turning it into a cartel. It might seem to offer the trade unions concerned the opportunity to push up their pay as far as the elasticity of demand for the products of the industry would let them. The difference between wages in the 'sheltered' and 'unsheltered' industries in the interwar years suggests that some effect of that kind did come about, if only in resisting downward pressure. More positive effects are less likely because employers' resistance will be based on their expectation of price rises stimulating competition from fresh sources at home as well as abroad.

Ideally, trade unions use establishment bargaining (the plant contract) to combine central control of 'the rate for the job' in all establishments, with whatever extra benefits can be extracted from the profitability of particular firms; but in hard times the local or union branch may prefer job security to maintenance of rates, and make concessions. Whereas industry-wide agreements are limited to simple provisions capable of general application, the American plant contract is generally voluminous, and provides rules for all manner of working practices and procedures in the plant. The trade union can therefore undertake to submit any dispute arising during the currency of the

contract to arbitration, as the arbitrator can interpret and apply the relevant rule to the facts of the case.

Trade Unionism at The Place of Work

Whether or not the trade unionists working in an establishment negotiate their own agreement with management, they are concerned with issues arising within its walls. Such issues include the allocation and pace of work; discipline; promotion; redundancies; and the processing of grievances. Under the law of the USA, the sole negotiating rights for all the manual workers of an establishment can be vested in one union; the officers of its local branch will then represent them on all these issues. In a British establishment the workers may belong to a number of unions, but the shop stewards elected by the members of the different unions come together in a council, which provides unified representation in meetings with management; its convenor may be wholly occupied with administrative business. Where the roots of trade unionism run back to handicrafts, the workshop is the arena in which the issues arise which bind the member to the union and over which sterner battles have been fought, as new machines and methods have come in, than have been caused by disputes about pay.

It has long been the aim of some trade unions in Europe, but not in the USA, to transcend the adversary system which opposed their members to management at the place of work. Many have sought to do this by a political revolution that would abolish capitalism, but equally in the social democracies, part of the case for nationalization has been that it would substitute public appointment for the irresponsibility of the private employer. Some trade unions have been more concerned with the substance of face-to-face relations, and the possibilities of workers' control and self-management. Interest has therefore attached to the statutory provision in the laws of some European countries, especially Germany, for works councils and the appointment of directors to represent the workers on supervisory boards. The general verdict on the German provisions is that the works councils – where the franchise extends to all employees, but the representatives are in practice the trade unionists – are greatly valued by the trade unions as a means of consultation and joint consideration of management issues; but the appointment of 'worker directors', though a mark of status whose removal would be resented, is not found to confer benefits that are actively felt.

The Impact of the Trade Union on Pay

Some estimate of the effects that trade unions have taken can be made by comparing the behaviour of pay in periods of trade union activity and at other times. In a number of Western countries there was a rapid extension of membership, for example, in the years following 1890; in Great Britain membership doubled during World War I; in many Western countries again, but not in the USA, membership rose in the years of full employment after World War II. When such indications as these of trade union strength and activity are set against the economic record, certain inferences suggest themselves about the extent to which trade unions may have changed the course of events, at large and in detail.

It appears that their effect on the general level of money wage rates has been in part to reinforce the ratchet effect which stops those rates dropping back and which has long been present even in the absence of combination: the much smaller reductions of wages in the organized trades in the USA in the great depression of 1929–34 is particularly striking. Generally it was observed that when the falling phase of the eight-year cycle brought wage cuts, the trade union deferred them or even staved them off altogether. Correspondingly, in the rising phase trade unionists were able to get a rise earlier than unorganized workers in their place would have done. But it has not appeared that even widespread and solid trade unionism has been able to push up the general level of money wage rates in a hard market environment, that is, when employers generally have not been able to pass

higher costs on in higher prices. The case has been different when the expectation of the employers, reinforced by the commitments of government, allow the negotiation of wage rises needed to keep the workers concerned in line with others, even though product prices must be raised in consequence: in these conditions associated with full employment the trade unions decide the course of the price level jointly with that of money wage rates.

The effect of trade unions on the level of real wages depends on their effect in the first place on productivity, or output per head, and then on their effect on distribution, or the share of output that accrues to the worker. That the 'restrictive practices' enforced by those unions whose control of employment is close enough serve to reduce productivity is evident from their nature, and from the willingness of managers to pay for their removal; but there are understandings about stints and working practices among unorganized workers too, and management must accept some understanding about these issues in any negotiated agreement with its workforce – the question is where the line shall be drawn. If changes in the strength of trade unionism have affected changes in productivity over time, it has been as only one among other and stronger influences: though the activity and spirit of the New Unionism in Great Britain were held responsible for the check to productivity that became conspicuous there at the beginning of the 20th century, the extension of trade union membership, and of trade union activity at the place of work in the 1950s and 1960s occurred at a time when the rise of productivity was exceptionally fast and sustained.

The effect of trade unions on distribution is illuminated by the evidence from a number of countries of trends that have kept real wages proportionate to productivity, that is, to real output per head (Phelps Brown and Browne 1968). Whatever the course of money wages, and whatever trade unions may have done at certain times to make them rise faster, the prices of products must have been adjusted so as to maintain a given ratio of wage to product; and in periods such as 1874–89 and 1923–37 in Great Britain, when money wages did not rise at all from end to end but productivity rose, the real wage was raised by a fall in prices. A further implication is that the proportionate division of the product between pay and profits has been constant. But this division, and the stability in the ratio of the real wage to productivity, has been subject to occasional displacement, in which that ratio has been raised. In depression and deflation, the power of the trade union to resist cuts compresses profit margins, and it appears that when the upheaval is sufficiently thoroughgoing, as after World War I, norms and expectations are permanently shifted, and the previous share of profits may never be restored.

Evidently the rise in the standard of living which has transformed the condition of the working population of the western world since 1850 seems to owe nothing directly to trade union pressure for higher wages. Trade unions appear to have taken more effect on distribution as anvil than as hammer. But these inferences from the behaviour of the general level of wages are compatible with substantial influence of the trade unions on the structure of pay. Particular groups may have gained by unionization. One effect of unionization is that it reduces the dispersion of rates for labour of the same type of grade, which otherwise is commonly wide, even in the same locality. Inquiries have also agreed in finding that unionization lifts the organized relatively to the unorganized. Collectively, this shows itself in the rise obtained when a group first bargains; but this is only an impact effect. There has been a cyclical pattern of variation between the wages of organized and unorganized workers, but not progressive divergence. Whether trade unions have changed the differential for skill depends on whether the skilled grades are organized and negotiate separately, or, if they belong to a general union, on their political influence within it. In Sweden in the 1950s and 1960s the pay structure was compressed by agreements made at the national level in pursuance of the egalitarian philosophy of the Landsorganisationen, the national trade union organization; but differentials were restored by wage drift on the shop floor. Statistical studies have shown that individuals who belong to trade unions earn substantially more than non-members when allowance is made for the

factors making up personal earning capacity: the difficulty is to be sure that all such factors have been taken into account.

Cost Push, Stagflation and Incomes Policies

The ability of trade unions to push up the general level of pay when employers are not constrained from raising prices became an engine of cost inflation in the 1960s, when trade unionists sloughed off the cautious expectations formed in harder times, and began raising their claims. Various forms of incomes policy were devised to persuade or require the trade unions to accept rises in money wages that did not outrun the prospective rise in productivity. But for the individual trade unionist, a rise in the money wage was equally a rise in the real wage at the time it was given; and experience showed that the tolerance of trade unionists for policies that required them to accept less than full compensation for rises in the cost of living, was limited. When in the 1970s recession brought back constraints on employers, the trade unionists' expectations and claims persisted, and the combination of unemployment and cost inflation was known as stagflation. It was widely recognized that in these circumstances an expansion of demand would be effective in reducing unemployment only if it was not used by trade unionists in jobs to push their pay up, and that it would therefore have to be accompanied by some form of agreement on restraint between the government and the trade unions.

See Also

- ▶ Arbitration
- ▶ Collective Bargaining
- ▶ Industrial Relations
- ▶ Strikes

Bibliography

Bain, G.S., and R. Price. 1960. *Profiles of union growth*. Oxford: Blackwell.

Hicks, J.R. 1932. *The theory of wages,* 2nd ed. London: Macmillan, 1963.

Phelps Brown, E.H., and M.H. Browne. 1968. *A century of pay: The course of pay and production in France, Germany, Sweden, the United Kingdom, and the United States of America, 1860–1960*. London: Macmillan.

Phelps Brown, E.H. and Hopkins, S.V. 1955. Seven centuries of building wages. *Economica* 22, August, 87. Reprinted in E.H. Phelps Brown and S.V. Hopkins, *A perspective of wages and prices*, London/New York: Methuen, 1981.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Ed. E. Cannan. Reprinted, London: Methuen, 1961.

Webb, S. and B. Webb. 1894. *The history of trade unionism,* 2nd ed. London: Longmans, 1920.

Trade, Technology Diffusion and Growth

Jonathan Eaton and Samuel Kortum

Abstract

Invention is a fundamental driver of growth in living standards around the world. Trade and technology diffusion are two ways the benefit of an invention spreads to other countries. Impediments to diffusion give rise to differences in living standards and to comparative advantage in production, providing an incentive to trade. We review some basic facts. We then provide a simple model of the connections among these processes, showing how it can explain these facts.

Keywords

Balanced growth; Capital goods trade; Cobb–Douglas functions; Diffusion of technology; Economic growth; Economic growth in the very long run; Endogenous growth; International trade; Invention; Patents; Product life cycle; Schmookler, J.; Technical change

JEL Classifications

D4; D10

'That the creation and diffusion of technological knowledge is at the heart of modern economic growth is now widely accepted' (Jacob Schmookler 1966). While technological advances drive world growth, differences in the ability of countries to exploit inventions create wide disparities in income. Understanding how such disparities arise requires identifying how the benefits of technological progress spread. A country might benefit from a new technique through trade, importing goods embodying that technique, or through diffusion, learning the technique and using it.

Impediments to technology diffusion mean that some countries have techniques that others lack, generating comparative advantage in production, as in Ricardo's model of trade. Over time, diffusion can eliminate or even reverse comparative advantage. As Vernon's (1966) product cycle posited, if diffusion is slow more inventive countries will export goods associated with recent inventions while importing goods embodying older technology. More recent work has modelled how technologies arise and spread across countries, driving world growth.

Before turning to how economists have modelled trade, diffusion and growth, we fill in this picture with some observations.

An Overview of the Evidence

When we look at trade in goods, several features stand out: (1) Over the second half of the 20th century, trade has grown as a share of GDP. (2) Nevertheless, geography continues to matter: countries buy much more from themselves and their neighbours than their shares in world production would imply. (3) Rich countries trade more among themselves and with poor countries than poor countries trade among each other.

Our ability to quantify invention is limited to observations on resources devoted to it (for example, research scientists and engineers) or on activities related to inventive output (for example, patents). These two measures paint a very similar picture: invention is concentrated in a small number of rich countries, although many rich countries are not particularly inventive. The patent data, like the trade data, have a further bilateral dimension as inventors from one country seek protection at home and elsewhere. As with merchandise trade, most patenting is done at home or nearby. Yet, most countries issue a majority of their patents to foreign inventors.

Income measures for the post-war era come from national accounts. Sifting through various pieces of evidence, economic historians, in particular Maddison (2003), have constructed measures going back several centuries. The basic picture here is that many countries have experienced sustained growth over a long period of time, with only infrequent switches in relative position.

Other evidence on invention and diffusion comes from longer ago. Diamond (1997) describes archaeological findings on when and where great innovations in agriculture, the domestication of plants and animals for farming, occurred, and how they diffused across continents. Sheep and wheat, for example, originated in south-western Asia around 8,500 BC and within a few millennia had diffused throughout the Eurasian land mass. Corn and turkeys had originated in Mesoamerica by 3,500 and slowly found their way through much of the Americas. Before Columbus, however, the two sets of technologies remained confined to their hemisphere of origin.

A Model

We now turn to a model of international technology diffusion, international trade and economic growth that is consistent with these observations.

A World of Ideas

We think about technology as a set of ideas that can be applied to production. New ideas generate technological advances if they lead to new goods and services or improve existing ones.

Ideas are inherently non-rival. An idea can be used in many places at the same time. Yet it may take a long time for an idea to spread. We sketch a simple formulation of this process, building on Nelson and Phelps (1966) and Krugman (1979).

We distinguish among three classes of a country's technology at any moment: (i) the measure of ideas T_i that country i invented itself, (ii) the measure of ideas T_i^A *available* to i, and (iii) the measure T_{ni}^E of *exclusive* ideas invented in i that are not yet available in n. Defining T^W as the set of ideas in the world:

$$T^W = \sum_i T_i.$$

New ideas arrive to country i at rate $\dot{T}_i$. Initially they are exclusive to i. They become known in country $n \neq i$, thus transiting from T_{ni}^E to T_n^A, with a hazard ε_{ni}.

Two sets of dynamic equations describe the evolution of technology. The first applies to the increase in available technology in country n:

$$\dot{T}_n^A = \dot{T}_n + \sum_{i \neq n} \varepsilon_{ni} T_{ni}^E.$$

The second applies to the change in its exclusive technologies:

$$\dot{T}_{ni}^E = \dot{T}_i + \sum_{n \neq i} \varepsilon_{ni} T_{ni}^E. \tag{1}$$

The literature has related invention to human effort and to research spillovers, acknowledging that inventors stand on the shoulders of others. A simple formulation highlighting research spillovers, used by Krugman (1979), is

$$\dot{T}_i = \iota_i T_i^A,$$

where $\iota_i \geq 0$ is an exogenous rate of invention in i. Romer (1990) models the determination of ι_i, turning it into an endogenous growth model. A simple formulation, highlighting human input, in the spirit of the semi-endogenous growth model of Jones (1995), is:

$$\dot{T}_i = \iota_i L_i.$$

Here L_i is the labour force in country i, which grows at a constant rate $g_L > 0$ in each country.

With an arbitrary number of countries, under general assumptions about the parameters of invention (the ι's) and diffusion (the ε's), patterns of the distribution of technology can be complex. Nevertheless, if these parameters remain constant, with the matrix of ε's indecomposable (ensuring that every proper subset of countries can absorb technology from outside), the world will evolve toward a balanced growth path with all T's growing at rate g_T. In the research spillover case g_T is given by an eigenvalue (the Frobenius root) of the matrix of ε's and ι's, and is increasing in both sets of parameters. In the human input case g_T equals the population growth rate g_L. In either case more inventive countries and those quicker to adopt technologies from elsewhere will have more ideas at any moment. Countries that are slow to adopt eventually fall far enough behind that they can draw from a stock of unknown technologies large enough to pull them along at the same growth rate as the most advanced (Eaton and Kortum 1999). Diamond's archaeological examples illustrate the phenomenon nicely.

A two-country example, using the human input specification of invention (so that $g_L = g_T > 0$) provides some basic results. With two countries we can simplify $T_{ni}^E = T_i^E$ and $\varepsilon_{ni} = \varepsilon_i$, and define the measure of ideas that have spilled out of either country's exclusive technology as T^C. It is straightforward to calculate the various technology measures along a balanced growth path:

$$\begin{aligned} T_i &= \frac{\iota_i}{g_L} L_i \quad T_i^E = \frac{g_L}{g_L + \varepsilon_i} T_i \quad TC \\ &= \frac{\varepsilon_1}{g_L + \varepsilon_1} T_1 + \frac{\varepsilon_2}{g_L + \varepsilon_2} T_2 \quad T_i^A \\ &= T_i + \frac{\varepsilon_n}{g_L + \varepsilon_n} T_n \quad n \neq 1. \end{aligned}$$

These equations show us how the various measures of technology, each growing at rate g_L, relate to the underlying parameters of invention and diffusion. In particular, a higher rate of diffusion ε_1 out of country 1 leads to more technology available in country 2.

For an idea to affect economic welfare it has to be connected with a good or service that people enjoy. We can then ask how the distribution of

ideas relates to production and trade, and thus how ideas ultimately affect welfare.

To make this connection we need to be concrete about preferences. We assume that utility can be represented as a constant-elasticity-of-substitution function of consumption over a wide variety of differentiated goods indexed by j. The implied price index is:

$$p = \left[\int_0^J p(j)^{-(\sigma-1)} dj\right]^{-1/(\sigma-1)}, \qquad (2)$$

where p(j) is the price of good j, J is the measure of goods, and $\sigma > 0$ is the elasticity parameter.

Ideas as New Goods

A particularly simple case equates an idea to a new good, with one worker required to produce one unit, as in Krugman (1979) and Romer (1990). We let $J \to \infty$ in (2) so as not to limit growth. A good not yet invented has an implicit price of infinity, hence to bound utility we require $\sigma > 1$.

Without trade countries can consume only the goods they know how to produce themselves. Under perfect competition a unit of any good that is produced will cost the local wage. Since T_i^A is the measure of goods produced in country i, the real wage in country i, which is the inverse of the price index there, is simply:

$$\frac{w_i}{p_i} = \left(T_i^A\right)^{1/(\sigma-1)}.$$

An increase in country i's available technology raises welfare by increasing the variety of available goods. More rapid technology diffusion out of country 1 helps country 2 while doing no harm to country 1. The dynamics of technology diffusion lead to parallel growth in living standards around the world, at rate $g_T/(\sigma - 1)$.

International trade allows a country to consume goods that it doesn't know how to make itself. In the two-country case with costless trade there are two cases to consider.

In the first, the countries' relative labour forces are in line with their access to technology. More precisely:

$$\frac{T_1^E}{T^W} < \frac{L_1}{L^W} < \frac{T_1^E + T^C}{T^W},$$

where $L^W = L_1 + L_2$. In this case $w_1 = w_2 = w$. Both countries will produce some goods using the common technology and all goods sell at a price equal to the common wage. The common real wage is:

$$\frac{W}{p} = \left(T^W\right)^{1/(\sigma-1)}.$$

Comparing this wage to the autarky wage above, trade perfectly substitutes for diffusion. Faster technology diffusion would require less trade, but leave welfare in both countries unchanged.

The other case emerges if one country, say country 1, is advanced technologically relative to the size of its labour force, or if:

$$\frac{T_1^E}{T^W} > \frac{L_1}{L^W}.$$

Now $w_1 > w_2$, and country 1 produces only goods associated with T_1^E, while 2 produces goods associated with all its available technology, which includes T_2^E and T^C. Country 1 exports its exclusive goods in exchange for both country 2's exclusive goods and the goods that both could in principle make. The demand for any good produced in country 1 relative to any good produced in country 2 is:

$$\frac{c_1}{c_2} = \left(\frac{w_1}{W_2}\right)^{-\sigma}$$

Since total demand for country 1 relative to country 2 labour is:

$$\frac{L_1}{L_2} = \frac{T_1^E c_1}{T_2^A c_2},$$

the relative wage satisfies:

$$\frac{w_1}{w_2} = \left(\frac{L_1}{L_2}\right)^{-1/\sigma} \left(\frac{T_1^E}{T_2^A}\right)^{1/\sigma}.$$

Country 1 has a higher standard of living the smaller it is and the *higher* the ratio of its exclusive ideas to the ideas available to country 2. The real wage in country 1 is

$$\frac{w_1}{p} = \left\{ T^W + \left[\left(\frac{w_1}{w_2} \right)^{\sigma - 1} - 1 \right] T_2^A \right\}^{1/(\sigma - 1)},$$

which exceeds $(T^W)^{1/(\sigma - 1)}$.

The two cases with international trade illustrate basic points about the relationship between technology diffusion, trade, and welfare. In the case of equal wages and costless trade the rate of diffusion does not matter for income or welfare. With unequal wages, more diffusion necessarily benefits country 2, while the effect on country 1 is ambiguous. If T_2^A starts off very small relative to T^W, an increase in diffusion benefits country 1 by allowing it to import more goods at a low price. But at higher values of T_2^A the gain from being able to import more goods at a low price is offset by the increase in the price it pays for all imports. At some point T_2^A gets so large that $w_1 = w_2$ and we are back in the previous case (bounding country 1's loss from 2's access to its technology).

In the equal wage case technology diffusion can act as a substitute for trade: once a good enters the common technology it may no longer be traded. In the second case, diffusion from country 1 to country 2 reverses the direction of trade. What 1 once exported it now imports.

A straightforward extension allows for multiple inventive, high-wage countries and multiple low-wage countries relying on the common technology. Innovative countries exchange goods associated with their exclusive technologies among themselves. They export these goods to low-wage countries in exchange for goods in the common technology. But low-wage countries have no incentive to trade goods in the common technology among themselves.

Ideas as Better Goods

Alternatively, an idea may not be a new good but rather a new technique for producing an existing good. This approach appears in Grossman and Helpman (1991), Aghion and Howitt (1992), Kortum (1997), and Eaton and Kortum (1999). Since there are no new goods, we set $J = 1$ in (2) and drop the restriction that the substitution elasticity exceeds 1.

Let's begin with a stark case of international trade, along the lines of Armington (1969), in which country 1 produces only goods $j \leq 1/2$ and country 2 only goods $j > 1/2$. The efficiency with which country i produces its range of goods is T_i. The demand for any good produced in country 1 relative to any good produced in country 2 is:

$$\frac{c_1}{c_2} = \left(\frac{w_1/T_1}{w_2/T_2} \right)^{-\sigma},$$

Since the total demand for country 1 relative to country 2 labour is:

$$\frac{L_1}{L_2} = \frac{c_1/T_1}{c_2/T_2},$$

the relative wage satisfies:

$$\frac{w_1}{w_2} = \left(\frac{L_1}{L_2} \right)^{-1/\sigma} \left(\frac{T_1}{T_2} \right)^{(\sigma - 1)/\sigma}.$$

As above, country 1's relative wage is higher the smaller it is, but its technological lead has an ambiguous effect, depending on σ.

Consider technological stagnation in country 2 so that $T_2 = 1$ for ever while T_1 grows at rate $g_T > 0$ (with relative labor forces not changing). If $\sigma < 1$ the stagnant country 2 actually grows faster while $\sigma > 1$ means faster growth for country 1. With Cobb–Douglas preferences ($\sigma = 1$) wages grow at the same rate: country 1's greater inventiveness is exactly offset by its worsening terms of trade. Oil-exporting countries, for example, can grow faster than the rest of the world even though they do not rank highly in standard technology indicators.

If we remove the Armington assumption, so that either country can in principle produce anything, the possibilities for parallel growth expand. Suppose that to produce a unit of any good in country 2 requires one worker, as does producing any good $j > 1/2$ in country 1. Country 1 can

produce any good $j \leq 1/2$ with efficiency $T_1 > 1$. If $\sigma < 1$, the stagnant country 2 cannot grow faster for ever. Once $w2$ hits w_1, thenceforth $w_2 = w_1$ as country 1 produces an ever greater range of goods $j > 1/2$, approaching its share of the world labour force asymptotically. In this case with inelastic demand, even with no technology diffusion trade spreads the benefits of 1's technological advances evenly. If $\sigma > 1$, on the other hand, then as country 1's technology improves it remains specialized in goods $j \leq 1/2$ and its relative wage grows for ever.

While this two-country example illustrates some basic points about innovation, diffusion and trade very neatly, it fails to account for a number of the basic facts discussed earlier. In particular, the world consists of many countries displaying varying degrees of innovative activity. While countries trade with each other, barriers to international trade remain substantial.

Eaton and Kortum (1999, 2002) develop a framework in which the same basic forces drive growth and trade among many heterogeneous countries separated by trade barriers. They treat an idea as a way to produce some good $j \in [0, 1]$ with some efficiency $q(j)$ drawn from a Pareto distribution with parameter θ. The implication is that, if country i has access to a measure T_i^A of ideas, its best technology for making good j, $z_i(j)$, is a realization drawn from:

$$G\left(z, T_i^A\right) = \Pr[Z_i \leq z] = e^{-T_i^A z^{-\theta}},$$

a type II extreme value distribution. Looking across the unit continuum of goods, $e^{-T_i z^{-\theta}}$ is the fraction that country i can produce with an efficiency no more than z.

With no international trade, the price of good j in country i is $p_i(j) = w/z_i(j)$. The real wage in country i as well as the standard of living there, is thus

$$\frac{w_i}{p_i} = \left[\int_0^\infty z^{\sigma-1} dG\left(z; T_i^A\right)\right]^{1/(\sigma-1)} = \gamma\left(T_i^A\right)^{1/\theta} i, \tag{3}$$

(where γ is constant that depends on θ and σ). Trade provides the potential to exploit comparative advantage. An iceberg transport costs d_{ni} separates country n from country i (meaning that delivering 1 unit of a good to country n requires shipping $d_{ni} \geq 1$ units from i). In the case of no diffusion $T_i^A = T_i^E$. The real wage in country I then becomes:

$$\frac{w_i}{p_i} = \gamma \frac{w_i}{\left[\sum_{k=1}^N T_k^E \left(w_k d_{ik}\right)^{-\theta}\right]^{-1/\theta}} \geq \gamma\left(T_i^E\right)^{1/\theta}.$$

Trade thus gives consumers in country i access to everyone's technologies, appropriately weighted by input and transportation costs.

With two countries we get the Ricardian model of Dornbusch et al. (1977). Ordering goods in decreasing order of country 1's relative efficiency $A(j) = z_1(j)/z_2(j)$ yields:

$$A(j) = \begin{cases} \left(\frac{1-j}{j}\frac{T_1^E}{T_2^A}\right)^{1/\theta} & j < \frac{T_1^E}{T^W} \\ 1 & \frac{T_1^E}{T^W} \leq j \leq \frac{T_1^A}{T^W} \\ \left(\frac{1-j}{j}\frac{T_1^A}{T_2^E}\right)^{1/\theta} & j > \frac{T_1^A}{T^W} \end{cases}.$$

Note that the range of goods over which $A(j)=1$ is governed by the extent of technology diffusion. With no transport costs and symmetric Cobb–Douglas preferences the equilibrium is characterized by a cutoff good j^* such that country 1 produces goods $j \in [0, j^*]$ while country 2 produces $j \in [j^*, 1]$. The relative wage w_1/w_2 and j^* satisfy the conditions for labour market clearing and comparative advantage:

$$\frac{w_1}{w_2} = \frac{L_2}{L_1}\frac{j^*}{1-j^*}\frac{w_1}{w_2} = A(j^*).$$

This two-country case is analyzed in Eaton and Kortum (2007a) with research effort determined endogenously. Rodriguez-Clare (2007) derives quantitative implications of an N-country case with research exogenous.

Ideas Embodied in Inputs

Our discussion so far has considered two ways in which the benefits of invention spread around the world: through trade in final goods that embody the technology, and through the diffusion of the disembodied idea itself. Ideas embodied in goods used for production provides a third conduit. Trade in capital goods is an important example. As Eaton and Kortum (2001) document, the production and export of capital goods is concentrated in a small number of research-intensive countries. For example, many countries have airlines that fly wide-bodied aircraft, but much of the technology for producing them is limited to Seattle and Toulouse.

Macroeconomists attribute a significant share of income growth to advances in capital equipment, causing its relative price to decline over time. Costless trade would imply the same relative price of capital goods everywhere, allowing all countries to benefit equally from inventions embodied in capital goods.

In fact, trade and price data suggest enormous geographic fragmentation of markets, with poor countries facing a systematically higher relative price of capital. A simple growth model translates differences in the relative price of capital goods to differences in per worker output. If capital has a share α in a Cobb–Douglas production function and depreciates at rate δ, country i with a savings rate s_i will have a steady-state level of output per worker y_{it} given by:

$$y_{it} = \left(\frac{s_i}{\left(\delta + \frac{g^K}{1-\alpha}\right)\left(P_{it}^K / P_i^C\right)} \right)^{\alpha/(1-\alpha)}.$$

Here P_{it}^K / P_i^C is the relative price of capital goods in country i at time t and g^K is the rate at which the price of capital goods declines. With constant iceberg trade barriers, g^K is the same everywhere in the world, but P_{it}^K / P_i^C is lower in countries with better access to capital goods. The formulation is thus consistent with a common world growth rate, but with persistent differences in levels over time.

Summary

We have sketched a simple model of trade, diffusion and growth motivated by some basic facts. A body of work has sought to quantify various models of this type.

Eaton and Kortum (2007b) review some of this work.

See Also

- Balanced Growth
- Comparative Advantage
- Diffusion of Technology
- Endogenous Growth Theory
- Growth and International Trade
- Inequality Between Nations
- Perron–Frobenius Theorem
- Product Life Cycle
- Ricardian Trade theory
- Schmookler, Jacob (1918–1967)
- Schumpeterian Growth and Growth Policy Design
- Technology
- Transfer of Technology

Bibliography

Aghion, P., and P. Howitt. 1992. A model of growth through creative destruction. *Econometrica* 60: 323–351.

Armington, P.S. 1969. A theory of demand for products distinguished by place of production. *IMF Staff Papers* 16: 159–178.

Diamond, J. 1997. *Guns, germs, and steel*. New York: W.W. Norton.

Dornbusch, R., S. Fischer, and P.A. Samuelson. 1977. Comparative advantage, trade, and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67: 823–839.

Eaton, J., and S. Kortum. 1999. International technology diffusion: Theory and measurement. *International Economic Review* 40: 537–570.

Eaton, J., and S. Kortum. 2001. Trade in capital goods. *European Economic Review* 45: 1195–1235.

Eaton, J., and S. Kortum. 2002. Technology, geography, and trade. *Econometrica* 70: 1741–1780.

Eaton, J., and Kortum, S. 2007a. Innovation, diffusion, and trade. In *Entrepreneurship, innovation, and the growth mechanism of the free-enterprise economies*, ed. E. Sheshinski, R. Strom and W. Baumol, 276–299. Princeton University Press.

Eaton, J., and S. Kortum. 2007b. *Technology and the global economy: A framework for quantitative analysis*. Manuscript: New York University and University of Chicago.
Grossman, G.M., and E. Helpman. 1991. *Innovation and growth in the global economy*. Cambridge, MA: MIT Press.
Jones, C.I. 1995. R&D-based models of economic growth. *Journal of Political Economy* 103: 759–784.
Kortum, S. 1997. Research, patenting, and technological change. *Econometrica* 65: 1389–1419.
Krugman, P.R. 1979. A model of innovation, technology transfer, and the world distribution of income. *Journal of Political Economy* 87: 253–266.
Maddison, A. 2003. *The world economy: Historical statistics (CD ROM)*. Paris: OECD.
Nelson, R.R., and E.S. Phelps. 1966. Investment in humans, technological diffusion, and economic growth. *American Economic Review* 56: 69–75.
Rodriguez-Clare, A. 2007. *Trade, diffusion, and the gains from openness*. Mimeo: Pennsylvania State University.
Romer, P.M. 1990. Endogenous technical change. *Journal of Political Economy* 98(5): S71–S102.
Schmookler, J. 1966. *Invention and economic growth*. Cambridge, MA: Harvard University Press.
Vernon, R. 1966. International investment and trade in the product cycle. *Quarterly Journal of Economics* 80: 190–207.

Tragedy of the Commons

Elinor Ostrom

Abstract

'The tragedy of the commons' arises when it is difficult and costly to exclude potential users from common-pool resources that yield finite flows of benefits, as a result of which those resources will be exhausted by rational, utility-maximizing individuals rather than conserved for the benefit of all. Pessimism about the possibility of users voluntarily cooperating to prevent overuse has led to widespread central control of common-pool resources. But such control has itself frequently resulted in resource overuse. In practice, especially where they can communicate, users often develop rules that limit resource use and conserve resources.

Keywords

Collective action; Common-pool resources; Common property; Free rider problem; Hardin, G.; Open-access resources; Private property; Social dilemmas; State property; Tragedy of the commons

JEL Classifications

D7; H23; Q2; Q21

The term 'the tragedy of the commons' was first introduced by Garrett Hardin (1968) in an important article in *Science*. Hardin asked us to envision a pasture 'open to all' in which each herder received large benefits from selling his or her own animals while facing only small costs of overgrazing. When the number of animals exceeds the capacity of the pasture, each herder is still motivated to add more animals since the herder receives all of the proceeds from the sale of animals and only a partial share of the cost of overgrazing. Hardin (1968, p. 1244) concluded:

> Therein is the tragedy. Each man is locked into a system that compels him to increase his herd without limit – in a world that is limited. Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons.

Hardin's article is one of the most cited publications of recent times as well as among the most influential for ecologists and environmental policy researchers. Almost all textbooks on environmental policy cite Hardin's article and discuss the problem that Hardin so graphically identified.

Hardin's article deals in general with a broad class of resources that are referred to in the more technical literature as 'common-pool resources'. Common-pool resources yield finite flows of benefits (such as firewood, fish and water) where it is difficult and costly to exclude potential users (Ostrom et al. 1994). Each person's use of a resource system subtracts resource units from the quantity of units available to others, as Hardin so dramatically described. The initial theoretical studies of common-pool resources tended to analyse simple systems. It has frequently been

assumed that the resource generates a predictable, finite supply of one type of resource unit (for example, cubic feet of water or tons of fish) in each time period. Users are assumed to be short-term, profit-maximizing actors who have complete information and are homogeneous in terms of their assets, skills, discount rates and cultural views. In this theory, *anyone* can enter a resource and take resource units.

Hardin thought of users as being trapped in this situation – largely because he did not envision that users could self-organize and devise institutions to extract themselves from tragic overuse. In the conventional textbook theory (Clark 1976), scholars have tended to agree with Hardin that the users could not extract themselves from this situation. Organizing so as to create rules that specify who is an authorized user and the rights and duties of authorized users creates a public good for those involved. All users benefit from this public good, whether they contribute or not (Olson 1965). Thus, getting 'out of the trap' is itself a second-level dilemma. Since much of the initial problem exists because the individuals are in a dilemma whereby they impose negative externalities on one another, it is not consistent with the conventional theory that individuals can solve a second-level dilemma when they are already predicted to be unable to solve the initial social dilemma. Thus, extensive free-riding is predicted in most efforts to self-organize and govern a resource as a community of users.

Because of these predictions and because many open-access resources have indeed resulted in tragic levels of overuse and sometimes destruction, many scholars and public officials have relied upon the conventional analysis to justify the need for centralized control of all common-pool resources. National legislation has been passed in many countries, and administrative responsibilities for managing natural resources have been turned over to centralized agencies. Unfortunately, the results of many of these efforts have been the opposite of what was hoped. Evidence has now been amassed that central regulation has frequently accelerated resource deterioration, complicated by several problems of corruption and inefficiency. In-depth case analyses have documented the accelerated overharvesting of forests that occurred after national governments declared themselves to be the owners of forested land (National Research Council 1986; Ascher 1995). Similar problems have occurred with inshore fisheries when national agencies presumed that they had exclusive jurisdiction over all coastal waters (Finlayson and McCay 1998).

Policy analysts tend to look for certainty and want to know whether the tragedy of the commons theory is either right or wrong. A more productive approach is to ask under what conditions it is correct and when it makes the wrong predictions. In settings where there is a large group, no one communicates, and where no rights to the resource exist, Hardin's theory is supported by considerable evidence. There are many settings in the world where the tragedy of the commons has occurred and continues to occur – ocean fisheries and the atmosphere being the most obvious.

Contrary to the conventional theory, however, multiple studies have demonstrated that users have overcome social dilemmas to craft institutions to govern their own resources (National Research Council 1986, 2002; McCay and Acheson 1987; Ostrom 1990, 2005). The possibility, however, that the users would find ways to organize themselves was not mentioned in basic economic textbooks on environmental problems until recently (compare Clark 1976, with Hackett 1998). The design principles that characterize robust, long-lasting, institutional arrangements for the governance of common-pool resources have been identified (Ostrom 1990) and supported by further testing (Guillet 1992; Morrow and Hull 1996; Weinstein 2000).

A recent National Research Council (2002) report provides an excellent overview of the substantial research showing that many common-pool resources are governed successfully by non-state provision units and that some government and private arrangements also succeed. No simple governance system has been shown to be successful in all settings (Dietz et al. 2003). Many of the robust resource governance systems documented in the above-cited research do not resemble the textbook versions of either a

government or a strictly private for-profit firm, especially when participants have constituted self-governing units. Scholars who draw on traditional conceptions of 'the market' and 'the state' have not recognized these self-organized systems as potentially viable forms of organization and have either called for their removal or ignored their existence. It is paradoxical that many vibrant, self-governed institutions have been wrongly classified or ignored in an era that many observers consider to be one of ever greater democratization.

Careful laboratory experiments have also shown that when a group of individuals are given unrestricted access to harvest from a common-pool resource, they substantially overuse it. What is rather striking is that in the laboratory using exactly the same parameters, but changing only one variable, namely, the capacity to communicate with one another, individuals can come to agreements and keep them to harvest very close to an optimal level (Ostrom et al. 1994). This result has been replicated many times (see, for example, Casari and Plott 2003).

Thus, Hardin opened a discourse on a fascinating and difficult puzzle of why individuals in some settings can overcome the threat to long-term sustainable use of a resource whereas other resources are so threatened. Scholars from multiple disciplines have wrestled with this question for several decades, including the creation of the International Association for the Study of Common Property (IASCP), the Scientific Committee on Problems of the Environment (SCOPE) (see Burger et al. 2001), considerable research in the field and in the experimental laboratory, and the development of sophisticated agent-based models of human-environmental relationships (Janssen 2003).

In the decades since Hardin's article appeared, we have learned that the type of resource must be analysed separately from the type of property arrangement. Common-pool resources exist wherever natural resources or human-made facilities exist and where excluding users is costly and consumption by some subtracts from the benefits available to others. Many types of property arrangements exist in relationship to these kinds of resources, including government ownership, private property and common property. Hardin incorrectly presumed that most common-pool resources were open-access resources where property rights had not been well-defined.

It is now known that the users of a common-pool resource will:

- expend considerable time and energy devising workable institutions for governing
- and managing common-pool resources;
- follow costly rules so long as they believe that others also follow these rules;
- monitor each other's conformity with these rules; and
- impose sanctions on each other at a cost to themselves.

The likelihood that resource users themselves will develop effective institutions for regulating the use of common-pool resources is increased by the following factors:

- low discount rates (most resource users have secure tenure, and plan on using the resource for a long time into the future);
- homogeneous interests (most resource users share similar technologies, skills, and cultural views of the resource);
- the cost of communication among individuals is low; and
- the cost of reaching binding and enforceable agreements is relatively low.

Thus, in field settings where there are relatively small-to moderate-sized groups, and where there is autonomy to make their own agreements and authority to do so, many user groups have self-organized to extract themselves from the tragedy.

Large groups have more difficulty governing common-pool resources, but usually because size is negatively associated with the factors listed above. In relatively homogeneous groups in which mechanisms exist for reaching binding agreements on methods of government and management resource use, even quite large groups are able to arrive at effective rules to limit the use of their resource. Further, when large groups are

composed of smaller groups that focus on specific parts of a larger problem, such as how to regulate water distribution on a branch of an irrigation canal, smaller groups can be clustered into ever larger aggregations that may be able to address problems that affect all participants.

One of the key findings of empirical field research on collective action and common-pool resources is the multiplicity of specific rules-in-use found in successful common-pool resource regimes around the world. One of the most important types of rules is *boundary* rules, which determine who has rights and responsibilities and what territory is covered by a particular governance unit. Many different boundary rules are used successfully to control common-pool resources around the world, but an important aspect of these rules is the match between the organization of users and the resource rather than the specific rule used. The 35th anniversary of the publication of Hardin's original article was celebrated with a special issue of *Science* (Dietz et al. 2003), demonstrating that all forms of ownership could succeed or fail and that more critical than the form of ownership was the establishment of legitimate and agreed-upon boundaries that were effectively enforced.

Some governance units face considerable biophysical constraints in dealing with a natural common-pool resource such as a groundwater basin, a river or an air shed. Such resources have their own geographic boundaries, and creating a match between the boundary of those who are authorized users and the resource itself is a challenge. On the other hand, the biophysical world does not have as strong an impact on the efficacy of using diverse boundaries for governing and managing forest resources. More important is the agreement of those involved about who is to be included and the appropriate physical boundaries. Rules specifying duties as well as rules for sharing benefits are also crucial. No resource system functions well over time if all that users do is harvest from it with no investment to increase the productivity of the resource itself. Once basic rules – defining who is a legitimate beneficiary, who must contribute to the maintenance of the resource, and the actions that must or may be taken or are forbidden – have been accepted as legitimate by the users, many users will follow rules so long as they believe others are doing so.

Another lesson learned is that any effort to develop new rules for governing and managing complex resources is likely to generate unexpected results and be subject to initial errors. Thus, all technological and institutional interventions need to be approached as an adaptive process that helps generate information about errors so that those involved and others can learn from errors rather than continue to make them. No panaceas exist. Wholesale solutions imposed on many different resources in a large terrain are more likely to be ineffective than efforts that enhance the institutional environment that encourages responsible self-governance, self-monitoring, and self-enforcement.

Thus, a modified theory of the commons is slowly evolving that has identified the factors that are repeatedly mentioned in empirical studies of diverse common-pool resources.

See Also

- ▶ Access to Land and Development
- ▶ Collective Action
- ▶ Common Property Resources
- ▶ Olson, Mancur (1932–1998)
- ▶ Property Law, Economics and
- ▶ Public Goods

Bibliography

Ascher, W. 1995. *Communities and sustainable forestry in developing countries*. San Francisco: ICS Press.

Burger, J., E. Ostrom, R. Norgaard, D. Policansky, and B. Goldstein. 2001. *Protecting the Commons: A framework for resource management in the Americas*. Washington, DC: Island Press.

Casari, M., and C. Plott. 2003. Decentralized management of common property resources: Experiments with a centuries-old institution. *Journal of Economic Behavior and Organization* 51: 217–247.

Clark, C. 1976. *Mathematical bioeconomics: The optimal management of renewable resources*. New York: Wiley.

Dietz, T., E. Ostrom, and P. Stern. 2003. The struggle to govern the commons. *Science* 302: 1907–1912.

Finlayson, A., and B.J. McCay. 1998. Crossing the threshold of ecosystem resilience: The commercial extinction

of northern cod. In *Linking social and ecological systems: Management practices and social mechanisms for building resilience*, ed. F. Berkes and C. Folke. New York: Cambridge University Press.
Guillet, D. 1992. *Covering ground: Communal water management and the state in the peruvian highlands*. Ann Arbor: University of Michigan Press.
Hackett, S. 1998. *Environmental and natural resources economics: Theory, policy, and the sustainable society.* Armonk: M. E. Sharpe.
Hardin, G. 1968. The tragedy of the commons. *Science* 162: 1243–1248.
Janssen, M., eds. 2003. *Complexity and ecosystem management: The theory and practice of multi-agent systems*. Northampton: Edward Elgar.
McCay, B., and J. Acheson. 1987. *The question of the commons: The culture and ecology of communal resources*. Tucson: University of Arizona Press.
Morrow, C., and R. Hull. 1996. Donor-initiated common pool resource institutions: The case of the Yanesha Forestry Cooperative. *World Development* 24: 1641–1657.
National Research Council. 1986. *Proceedings of the conference on common property resource management*. Washington, DC: National Academy Press.
National Research Council. 2002. The drama of the commons, Committee on the Human Dimensions of Global Change, ed. E. Ostrom, T. Dietz, N. Dolšak, P. Stern, S. Stonich and E. Weber. Washington, DC: National Academy Press.
Olson, M. 1965. *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.
Ostrom, E. 2005. *Understanding institutional diversity.* Princeton: Princeton University Press.
Ostrom, E., R. Gardner, and J. Walker. 1994. *Rules, games, and common-pool resources*. Ann Arbor: University of Michigan Press.
Weinstein, M. 2000. Pieces of the puzzle: Solutions for community-based fisheries management from native Canadians, Japanese cooperatives, and common property researchers. *Georgetown International Environmental Law Review* 12: 375–412.

Transaction Costs

Jürg Niehans

Transaction costs arise from the transfer of ownership or, more generally, of property rights. They are a concomitant of decentralized ownership rights, private property and exchange. In a collectivist economy with completely centralized decision-making they would be absent; administrative costs would take their place.

In modern economies a substantial, and probably increasing, proportion of resources is allocated to transaction costs. Nevertheless, up to World War II economic theory had virtually nothing to say about them. Over the last few decades a large and diverse literature has developed, but the analytic complexities are such that success still is only partial; important problems remain unsolved.

Transaction Technology

Transaction costs, like production costs, are a catch-all term for a heterogeneous assortment of inputs. The parties to a contract have to find each other, they have to communicate and to exchange information. The goods must be described, inspected, weighed and measured. Contracts are drawn up, lawyers may be consulted, title is transferred and records have to be kept. In some cases, compliance needs to be enforced through legal action and breach of contract may lead to litigation.

Transaction costs face the individual trader in two forms, namely (1) as inputs of his own resources, including time and (2) as margins between the buying and the selling price he finds for the same commodity in the market.

The transaction technology specifies what resource inputs are required to achieve a given transfer. It may be formalized in a ‘transaction function’ analogous to a production function. In principle, each such function relates to a specific pair (or, more generally, group) of economic agents. In this respect transaction costs are analytically analogous to transportation costs, which relate to a pair of locations. In one way or another, transaction costs are incurred in an effort to reduce uncertainty. For many purposes it may nevertheless be an efficient research strategy to proceed *as if* transaction costs occurred even under full certainty. Transaction costs then become, as Stigler (1967) put it, ‘the costs of transportation from ignorance to omniscience’.

While transaction costs are analogous to transportation costs in some respects, they are quite different in others. This is because they relate not to individual commodity flows, but to pairs (or, more generally, to groups) of such flows. There must be a *quid pro quo* in every single transaction. This requirement imposes constraints for which there is no spatial counterpart. While in a Walrasian equilibrium each trader has to observe only his budget constraint, in a transaction cost equilibrium he has to balance his account with every other trader. This gives rise to an additional set of shadow prices, reflecting the burden of the bilateral balance requirement (Niehans 1969).

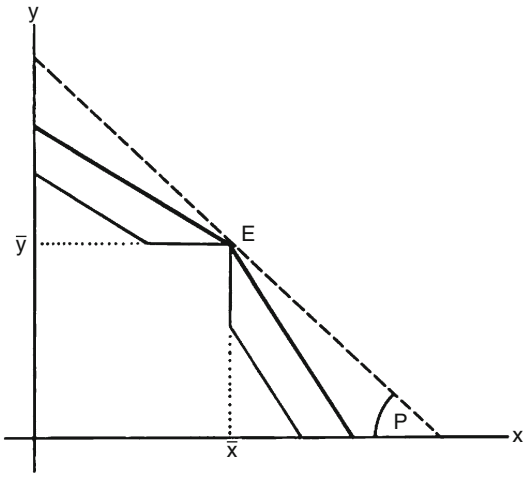
Transaction functions may exhibit diminishing, constant, or increasing returns. Scale economies are often pronounced; in many cases, transaction costs are virtually independent of the quantity transferred. The scale effects may relate to the size of the individual transaction, to the size of the participating firm or to the size of the market as a whole.

Only for simple exchange will a transaction function, built in analogy to a production function, provide an adequate description of transaction technology. Many contracts, particularly the more important ones, are far more complicated, often assuming a bewildering (and expensive) complexity. As a consequence, transaction costs become difficult, and perhaps impossible, to quantify. The analysis of more complex contracts, institutions and economic arrangements has thus been forced to rely more on qualitative than on quantitative methods.

The Volume of Transactions

Transaction costs, by and large, reduce the volume of transactions. In general equilibrium without transaction costs, the network of exchanges is indeterminate; there is no constraint on the gross trading volume. With increasingly costly transactions, individuals have an ever stronger incentive to economize transactions.

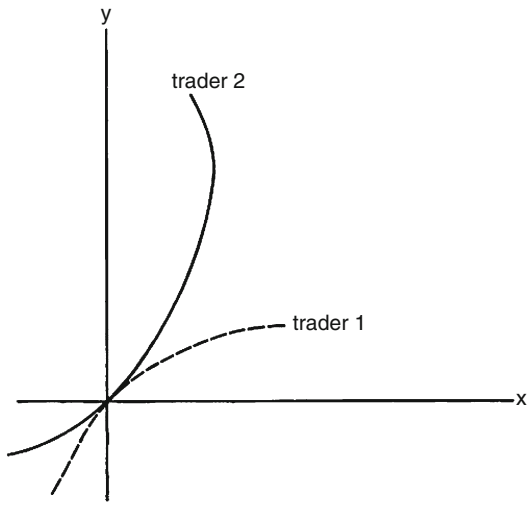
This can be clearly seen in a single market in which x is exchanged against y. In the budget constraint confronting an individual trader,



Transaction Costs, Fig. 1

proportional transaction costs produce a kink. If they amount to θ units of y for every unit of x bought or sold, the constraint will look like the heavy line in Fig. 1, where $\bar{x}$ and $\bar{y}$ mark the initial endowment. Depending on the shape of the indifference curves, the individual may wish to buy x (selling y), to buy y (selling x), or not to trade at all. A shift in the market price will let the budget constraint swivel around E. The important point is that, because of the kink, there is a *range* of prices for which trade remains zero.

The reciprocal demand curves for two representative traders will, with transaction costs, have a kink at the origin (Fig. 2). This may have the consequence that trade remains at zero (as illustrated) despite considerable changes in tastes and/or endowments. If transaction costs also have a fixed component, the budget constraint assumes the shape of the thin line in Fig. 1, and the reciprocal demand curves have an empty space around the origin (not illustrated). The larger transaction costs, both variable and fixed, the more likely it is that equilibrium is at the no-trade point. In a multimarket framework, therefore, transaction costs can explain why certain potential markets, either for present or future goods, do not exist (Niehans 1971). There have been numerous studies applying these general considerations to particular markets.



Transaction Costs, Fig. 2

One way of economizing on costly market transactions is the establishment of firms. Coase (1937) regarded the cost of using the price mechanism as the main reason for the existence of firms. For Williamson (1979, 1981) transaction costs are not only the key to an institutional theory of the firm but also to a new type of institutional economics.

The Bunching of Transactions

Fixed transaction costs tend to result in a bunching of transactions. This effect has played a major role in explaining the demand for money. Cash balances are held because for short holding periods the costs of buying and selling an earning asset are too high compared to its yield (Hicks 1935). Using an elementary inventory model with a sawtooth pattern of total assets, Baumol (1952) and Tobin (1956) derived algebraic demand functions for the demand for money. With fixed transaction costs, the demand for money would rise only with the square root of total assets, but this property does not hold in more general models. There is a vast literature applying the Baumol/Tobin approach to problems of monetary economics. In the history of economic thought few quantitative models of comparable simplicity have inspired more widespread uses.

Efficiency

Compared to an imaginary state with costless transactions, transaction costs inevitably reduce welfare. In the individual optimization model of above, the set of consumption possibilities shrinks. The welfare loss is reflected partly in the resources allocated to transactions and partly in the suppression of exchanges that would otherwise have been mutually beneficial.

The more interesting question is whether transaction costs make an economy inefficient. A number of contributions to it are surveyed by Ulph and Ulph (1975). The mere fact that the Walrasian auctioneer uses up resources, reflected in a spread between selling and buying prices, does not in itself create efficiency problems. However, increasing returns in transaction technology, particularly in the form of fixed transaction costs, may lead to distortions. It is well known that in the presence of scale economies competition may not lead to an efficient allocation of resources.

Hahn (1971) took the view that transaction costs generally result in an inefficient equilibrium because the multiplicity of budget constraints reduce consumption possibilities. Kurz (1974a, b) made it clear, however, that the alleged inefficiency may just be due to an inappropriate efficiency concept. The real question is whether, with given initial allocation and given transaction technology, the resulting equilibrium could be improved upon by a Paretosuperior reallocation, even though this would again cost resources. In the absence of scale economies, the discussion has produced no reason why, in this sense, transaction costs should generally cause inefficiency.

Efficiency problems also arise in a more general context. Simple exchange is a bilateral transaction. More complicated transactions may range from triangular exchange to multilateral contracts with a large number of parties. With increasing complexity, transaction costs tend to increase very rapidly. Even triangular contracts, therefore, are relatively rare and for more complex transactions the costs may rapidly become prohibitive. This is the basic reason for the emergence of market economies consisting of a network of bilateral exchanges. Politics may be interpreted as the

arena in which multilateral transactions are typically made.

In a sense, any deviations from Pareto-optimality can be attributed to transaction costs, because in their absence all opportunities for Pareto-superior contracts would be realized. This is the so-called 'Coase theorem' (Coase 1960). If, for example, the externalities of water pollution give rise to a social loss, one can imagine a multilateral abatement contract providing for payments from the sufferers to the polluters which is beneficial to all. In a world without transaction costs, therefore, private contracts could take the place of regulation. In the real world, however, as Coase emphasized, multilateral contracts tend to be very costly. Regulation, therefore, may be efficient, not because there is an externality, but because regulation may be cheaper than a multilateral contract. A similar reasoning can be applied to monopoly (Demsetz 1968).

Buchanan and Tullock (1962) have extended this type of analysis to political decisions, where the individual is assumed to weigh his benefit from collective action against his share of decision-making costs. If the latter are zero for everybody, a unanimity rule would lead to a Pareto optimum, but in the presence of transaction costs the high costs of unanimity are likely to result in other decision rules. In the debate about these propositions it has often been pointed out that the underlying definition of transaction costs may be tautological: Whatever produces deviations from Pareto-optimality is implicitly interpreted as a transaction cost.

Arbitrage

Transaction costs, like transportation costs, obstruct arbitrage, thus impeding the Law of One Price. Suppose, in an efficient and competitive exchange network, goods, on their way from producers to consumers, pass through the hands of several middlemen. Along each link of the network the increase in price will just pay for the marginal transaction costs. Where transaction costs would exceed the price differential, the transaction does not take place; in the reverse case the shortfall will be eliminated by competition. If between two potential intermediaries no transactions take place, their prices, within the margin of transaction costs, may fluctuate independently without calling forth a commodity flow. The one market price is thus replaced by a cluster of prices. Markets with transaction costs are often called 'imperfect'. This should not be regarded as a value judgement. In the presence of transaction costs, efficiency requires a multiplicity of prices.

Transaction costs also limit arbitrage between different assets. In a multicommodity exchange system in which every good can be exchanged against each of the others, perfect markets would result in consistent 'cross rates'. The foreign-exchange market is a good example: in the absence of transaction costs, the sterling rate of the dollar equals the sterling rate of the mark times the mark rate of the dollar. With transaction costs, the equality is replaced by a set of inequalities.

The influence of transaction costs on asset arbitrage was studied for many particular markets, including those for Eurocurrencies (Frenkel and Levich 1975), bonds of different types (Litzenberger and Rolfo 1984) and maturities (Malkiel 1966), stocks (Demsetz 1968, is the forerunner of many studies), take-overs (Smiley 1976), stock options (Phillips and Smith 1980) and commodities (Protopapadakis and Stoll 1983).

Intermediation

Imagine an economy in which all exchange consists of bilateral barter. In the absence of transaction costs it would make no difference who trades with whom; on their way from producers to consumers, commodities could pass through any number of hands. The presence of transaction costs makes the exchange network determinate. In such a network, certain traders, in view of their lower transaction costs, probably emerge as middlemen, brokers or intermediaries (Niehans 1969). A pure intermediary makes his contribution to the social product, abstracting from any associated contribution to production, by helping other traders to economize on transaction costs.

Transaction costs, therefore, are the key to an understanding of intermediation and of the structure of markets.

This is especially important in asset markets. For many consumer goods, particularly perishable ones, transaction costs are too high for them to pass through many hands. However, for assets like deposits, securities, foreign exchange, commodity contracts, gold options, insurance contracts, and mortgages, transaction costs are low enough to permit complicated intermediary networks. Benston and Smith (1976) thus argued convincingly that transaction costs are the *raison d'être* of financial intermediaries.

The Eurodollar market offers an instructive example. In the interwar period it became customary to regard banks primarily as producers of money and possibly other liquid assets. From this point of view, the emergence and the functioning of the Eurodollar market appeared as a 'puzzle'. The puzzle was easily solved once it was realized that the market for dollar funds (and other currencies) tended to move wherever transaction costs were lowest (Niehans and Hewson 1976). The more transaction costs decline under the pressure of financial innovation, the more highly developed will be the division of labour in financial services, the more elaborate the structure of the financial system and the higher the flow of daily transactions compared to the stocks of traded assets. It is tempting, therefore, to interpret the rapid changes in financial markets in recent years largely as a consequence of changing transaction costs.

Media of Exchange

Transaction costs are also responsible for the use and choice of media of exchange. The lower transaction costs on a given commodity, the more likely that this commodity will serve as money. Thanks to low transaction and holding costs, money helps to save resources that would otherwise have been used up in transactions. More important, it extends the scope of mutually beneficial exchange. In a world with transaction (and holding) costs, money thus has (indirect) utility even though, being a mere token money, it may have no direct utility.

Though this insight is old, its analytical implementation has made progress only in the last two decades. A simple expedient is to express transaction costs as a declining function of cash balances and then treat them like other costs (Saving 1971, 1972), but this begs the question how exactly such a function is determined.

The services of money for the individual consumer in the presence of transaction costs were analysed by Bernholz (1965, 1967) and, more fully, by Karni (1973). A rigorous analysis would have to be based on a general-equilibrium model of bilateral barter with transaction costs, which is not yet available. Since cash balances are an inventory, this needs to be a multiperiod model in which endowments, tastes and perhaps technology are subject to fluctuations. In order to model such fluctuations in an equilibrium framework, one might visualize those changes in the form of infinite stationary motion in which successive 'days' or 'seasons' are different, but successive 'years' are the same.

Such an economy will generally exhibit a complex pattern of markets in which a given commodity is traded against many (though not all) other commodities. If, from this arbitrary starting point, transaction costs are gradually lowered for one particular good, this good appears as the *quid pro quo* in an increasing number of transactions, while other barter exchanges disappear. There may also be cases with several moneys, each with its comparative advantages (Niehans 1969). If transaction costs on the medium of exchange (and also its holding costs) are low enough, it will be used as a general medium of exchange. If, in the limit, money can be transferred, produced and held without cost, one arrives at the special case of a Walrasian economy with an integrated budget constraint and neutral money (Niehans 1971, 1975, 1978), but, in contrast to Walras, with a determinate exchange network.

The rigorous mathematical analysis of the existence, uniqueness and efficiency of monetary equilibria with transaction costs made some progress during the 1970s (see Honkapohja 1977,

1978a, b and the literature given there). Since then, progress has been slow. The difficult process of adapting the traditional concepts of general-equilibrium analysis to the requirements of an intertemporal transaction-cost economy is still incomplete. This is one area where rigour so far has been at the expense of substance.

See Also

- ▶ Economic Organization and Transaction Costs
- ▶ Money and General Equilibrium Theory

Bibliography

Baumol, W.J. 1952. The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics* 66(4): 545–556.

Benston, G.J., and C.W. Smith. 1976. A transactions cost approach to the theory of financial intermediation. *Journal of Finance* 31(2): 215–231.

Bernholz, P. 1965. Aufbewahrungs- und Transportkosten als Bestimmungsgründe der Geldnachfrage. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 101(1): 1–15.

Bernholz, P. 1967. Erwerbskosten, Laufzeit und Charakter zinstragender Forderungen als Bestimmungsgründe der Geldnachfrage der Haushalte. *Zeitschrift für die gesamte Staatswissenschaft* 123(1): 9–24.

Buchanan, J.M., and G. Tullock. 1962. *The calculus of consent; Logical foundations of constitutional democracy.* Ann Arbor: University of Michigan Press.

Coase, R.H. 1937. The nature of the firm. *Economica* 4(16): 386–405.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Demsetz, H. 1968. The cost of transacting. *Quarterly Journal of Economics* 82(1): 33–53.

Frenkel, J.A., and R.M. Levich. 1975. Covered interest arbitrage: Unexploited profits? *Journal of Political Economy* 83(2): 325–338.

Hahn, F.H. 1971. Equilibrium with transaction costs. *Econometrica* 39(3): 417–439.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* 2(1): 1–19.

Honkapohja, S. 1977. Money and the core in a sequence economy with transaction costs. *European Economic Review* 10(2): 241–251.

Honkapohja, S. 1978a. A reexamination of the store of value in a sequence economy with transaction costs. *Journal of Economic Theory* 18(2): 278–293.

Honkapohja, S. 1978b. On the efficiency of a competitive monetary equilibrium with transaction costs. *Review of Economic Studies* 45(3): 405–415.

Karni, E. 1973. Transactions costs and the demand for media of exchange. *Western Economic Journal* 11(1): 71–80.

Kurz, M. 1974a. Equilibrium in a finite sequence of markets with transactions cost. *Econometrica* 42(1): 1–20.

Kurz, M. 1974b. Arrow-Debreu equilibrium of an exchange economy with transaction cost. *International Economic Review* 15(3): 699–717.

Litzenberger, R.H., and J. Rolfo. 1984. Arbitrage pricing, transaction costs and taxation of capital gains: A study of government bonds with the same maturity date. *Journal of Financial Economics* 13: 337–351.

Malkiel, B.G. 1966. *The term structure of interest rates: Expectations and behavior patterns*. Princeton: Princeton University Press.

Niehans, J. 1969. Money in a static theory of optimal payment arrangements. *Journal of Money, Credit and Banking* 1(4): 706–726.

Niehans, J. 1971. Money and barter in general equilibrium with transactions costs. *American Economic Review* 61(5): 773–783.

Niehans, J. 1975. Interest and credit in general equilibrium with transaction costs. *American Economic Review* 65(4): 548–566.

Niehans, J. 1978. *The theory of money.* Baltimore: Johns Hopkins University Press.

Niehans, J., and J. Hewson. 1976. The eurodollar market and monetary theory. *Journal of Money, Credit and Banking* 7(1): 1–27.

Phillips, S.M., and C.W. Smith. 1980. Trading costs for listed options: The implications for market efficiency. *Journal of Financial Economics* 8: 179–201.

Protopapadakis, A., and H.R. Stoll. 1983. Spot and futures prices and the law of one price. *Journal of Finance* 38(5): 1431–1455.

Saving, T.R. 1971. Transactions costs and the demand for money. *American Economic Review* 61(3): 407–420.

Saving, T.R. 1972. Transactions costs and the firm's demand for money. *Journal of Money, Credit and Banking* 4(2): 245–259.

Smiley, R. 1976. Tender offers, transactions costs and the theory of the firm. *Review of Economics and Statistics* 58(1): 22–32.

Stigler, G.J. 1967. Imperfections in the capital market. *Journal of Political Economy* 75(3): 287–292.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38(3): 241–247.

Ulph, A.M., and D.T. Ulph. 1975. Transaction costs in general equilibrium theory: A survey. *Economica* 42(168): 355–372.

Williamson, O.E. 1979. Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics* 22(2): 233–261.

Williamson, O.E. 1981. The modern corporation: Origins, evolution, attributes. *Journal of Economic Literature* 19(4): 1537–1568.

Transaction Costs, History Of

M. Klaes

Abstract

While the basic insight that underlies the transaction cost concept is probably as old as human reflection on economic issues itself, it became associated in the 19th century with the notion of economic friction, which was subsequently expressed as a cost. Historically, the transaction cost concept has developed from narrow interpretations typical of the monetary and general equilibrium literature towards relational interpretations, based on particular market microstructural models of how economic agents interact with each other, and finally with institutional interpretations embracing a more general analysis of economic institutions, including market and non-market forms of coordination.

Keywords

Cash balances; Circulation cost; Coase, R. H; Contract theory; Coordination problem; Firm, theory of; Frictions; Institutional economics; Institutional transaction cost economics; Law and economics; Link costs; Modularity; Money; New institutional economics; Non-market coordination; Organization, economics of; Productive and unproductive labour; Property rights; Transaction cost economics; Transactions costs; Transactions demand for money; Marschak, J; Williamson, O. E

JEL Classifications

B1

As a concept, transaction costs are used in numerous ways in economics, from simply referring to the fees charged by a financial broker to a much broader concept encompassing the comparative efficiency of alternative modes of resource allocation and economic coordination.

At the most general level, transaction costs are the costs that arise beyond the point of production of a good to effect its allocation. From there on, the literature is fragmented regarding the various facets of the concept. The distinction between production and allocation may not be meaningful in all instances. Transaction costs may or may not include transport costs, may or may not refer only to market exchange, may or may not be reduced to a single alternative category such as information costs or the cost of time. Some authors measure transaction costs in monetary terms, others as departures from first-best outcomes, or just on the basis of qualitative comparative rankings of feasible institutional alternatives. Indeed, whether transaction costs should be regarded as costs at all has been subject to controversy, too. Hence, the term can be and has been applied in virtually every conceivable economic and social scientific context. Its wide diffusion has been attributed to the systematic ambiguity inherent in its unqualified application (Klaes 2001), and its usefulness for serious analysis has been questioned on these grounds (for example, Dixit 1996, p. 35).

Although often used as a catch-all expression, thinking in transaction cost terms has nevertheless yielded a rich array of models and analytical frameworks that have helped redefine how economists look at economic exchange and coordination. Circumspect definition specific to the particular context in which one seeks to use the concept should help to avoid semantic pitfalls. Furthermore, not only are systematically ambiguous notions common in economics (Clower 1995), they actually serve a distinct and valuable purpose in the coordination of research (Klaes 2006). A sceptical stance towards the aggregate ambiguity of the transaction cost concept does thus not necessarily amount to a criticism of particular applications of the term, although a certain tendency can be observed outside of what has become known as a ‘new’ institutional economics (see below) towards alternative expressions such as ‘friction’ (for example, Luttmer 1996) or ‘link costs’ (for example, Kranton and Minehart 2001, p. 492).

Monetary, Relational and Institutional Interpretations

The various interpretations of the transaction cost concept have traditionally been classified into two general categories, juxtaposing a narrow interpretation typically associated and compatible with a 'neoclassical' tradition, and a broader and more institutionally minded interpretation, located in the theory of the firm and the economics of property rights, and calling for more or less radical revisions of this tradition (Coase 1972; cf. Dahlman 1979; Allen 2000). While careful analysis allows one to distinguish between more than two categorically different kinds of transaction cost, the range of extant applications of the concept is best thought of as forming a spectrum of broadening scope: (a) narrow interpretations typical of the monetary and general equilibrium literature; (b) relational interpretations that are based on particular market micro-structural conceptions or models of how economic agents interact with each other beyond the traditional economic dimensions of price and quantity signals; (c) institutional interpretations, which formulate transaction costs as part of a more general analysis of economic institutions, including market and non-market forms of coordination. Institutional interpretations of transaction costs are the result of applying relational interpretations of transaction costs – originally defined on the basis of exchange within markets – to non-market settings such as networks, firms or clans, with the aim of expressing the comparative economic performance of alternative institutional solutions to the coordination problem in transaction cost terms.

In monetary conceptions, transaction costs are the direct costs that an economic agent incurs when engaging in a market transaction, if we leave most or all of the microstructural details of the exchange context unspecified. At the most basic level, these costs are expressed as a reduction in the value of a transaction, technically equivalent to a transaction tax. In more developed interpretations of monetary transaction costs, these costs are conceptualized as the direct monetary costs incurred when engaging in a particular market transaction, resulting from the use of intermediary and adjunct services (brokerage, transport).

Relational transaction cost interpretations rely on a more explicit conceptualization than monetary transaction cost interpretations of how economic agents interact with one another when they engage in market exchange. To some extent, one may subsume economic contract theory under this part of the transaction cost literature, although explicit transaction cost conceptions as such play at best a subordinate role in those approaches (for example, Grossman and Hart 1986; Holmstrom and Milgrom 1994).

A relational interpretation of transaction costs has played a more pronounced role in Coase's (1937, 1960) contributions to the theory of the firm and the economics of property rights. While he himself referred to 'marketing costs' or the 'costs of market transactions' in those seminal papers and did not embrace the term 'transaction costs' until relatively late (Coase 1974, p. 494), his approach to defining these costs heavily influenced the conceptualization of transaction costs in what has become known as a 'new' institutional economics (Eggertsson 1990; Furubotn and Richter 2005; Ménard and Shirley 2005), a movement best thought of as part of a renewed interest in the institutionalist traditions in economics during the last decades of the 20th century (Rutherford 1994). Faced with the considerable theoretical challenge of providing a comprehensive micro-structural theory of exchange, a frequent strategy in this literature has been to follow Coase in decomposing relational transaction costs heuristically according to the different steps involved in concluding a market transaction. While classifications of individual authors differ, most heuristics can be accommodated within a framework that distinguishes between: (a) the costs of locating and attracting potential trading partners and of presale inspection; (b) contracting and fulfilment costs; (c) policing and enforcement costs.

In most instances, relational conceptions of transaction costs still proceed from a contractual (in the legal sense) and therefore market-based understanding of 'transaction'. Institutional interpretations, by contrast, further broadened the scope of the transaction cost notion by applying it not just

to contractual settings but also to alternative forms of economic coordination. The distinguishing characteristic of this third interpretation of transaction costs is not the comparative character of the underlying analysis, since both monetary and relational conceptions allow comparative assessments of alternative solutions to a given coordination problem. Rather, the difference results from the endeavour of applying Coase's 'marketing' costs to non-market settings, comparing market coordination alongside non-market forms within a given array of alternative institutional forms. However they are defined on the micro-structural level, once they are institutionally interpreted transaction costs therefore reflect the costs of economic coordination more generally.

Some economists regard the transaction cost concept as inseparably wedded to a framework of analysis that attempts to reduce institutional features of the economy to the core neoclassical notion of cost, thereby failing to develop conceptual tools more attuned to the complexities of economic institutions. In the eyes of these critics, building an analysis of the institutional features of economic coordination on a concept that invites its own minimization undermines the very starting point of any seriously institutional approach to economics, since a world of zero transaction costs would constitute an institutional void. In other words, one may argue that an important aspect of transaction costs is their reflection of investment in institutional capital. In a way, this point mirrors the various critiques of reducing labour to an economic cost. It has had limited impact on those who have sought to reform and expand economic analysis in the name of the economics of property rights, the field of law and economics, transaction cost economics, the theory of the firm, the economics of organization, and the study of long-term economic change (see Coase 1988; Williamson 2000; Alchian 1991; Demsetz 1988; North 1990; Langlois 2002).

Origins

The basic insight that underlies the transaction cost concept, which one finds embodied in pre-historical emergence of generally accepted media of exchange, is probably as old as human reflection on economic issues itself. It is thus not difficult to identify numerous precursors to the notion of transaction costs as one finds it presently developed in the economic literature, in particular in its monetary branches. When Aristotle (1932, pp. 13–14) wrote on the origin of money for example, he observed that once villages grew and combined into city-states they would require a medium of exchange that was portable and easy to handle. He also noted that impressing a stamp on a piece of metal would help avoiding repeat measurement of its embodied value.

Aristotle's insights have been reiterated time and again whenever subsequent economic writers discussed the origin of money. Crucial steps for the further entrenchment of the concept were the depiction of the various impediments to exchange as an economic cost, followed by the crystallization of the term 'transaction(s) costs' itself (probably Scitovsky 1940, p. 307; cf. Hardt 2006), which entered the economic literature from the financial markets where it was in popular usage in the 1930s (Anon 1936).

Towards the end of the 19th century, economists tended to address the impediments to exchange as 'frictions' in the economic system (cf. Davidson 1896). On the metaphorical level, this interpretation resulted from a general post-Enlightenment tendency to look at the economic system in mechanistic terms, illustrated by recurring metaphors such as Hume's (1752, II, iii, 1) 'wheels of trade' or Mill's (1848, III, xxvi, 1) 'machinery of exchange'. One should not discount the economic cost of the physical friction that a medium of exchange is exposed to either (Say 1803, I.XXI.xi, 1–3).

A prominent early attempt to describe economic friction as a cost can be found in Menger (1871, pp. 170–1), who notes that every transaction requires economic sacrifices. At the very minimum, these consist of a loss of time, but may also include transport and storage costs, sales costs, taxes, commissions, communication costs, and more generally all the costs associated with intermediaries and the monetary system. While clearly describing these various sacrifices

in opportunity-cost terms, Menger does not address them explicitly as a cost and he refers to them merely to define the boundary conditions of the validity of his theory of exchange.

It was not until Marx's (1893, pp. 123–46) analysis of Zirkulationskosten, as the 'costs of circulation' that result from the continuous exchange of money into goods and back into money again, that one finds an extended conceptual analysis of the costs associated with exchange. Marx regarded these costs as 'faux frais', costs which are necessary to sustain the circulation of capital but are nevertheless unproductive in that they do not contribute to the creation of value. 'Pure circulation costs', the most important component of circulation cost, refers to bargaining costs, accounting costs, and, following in Say's footsteps, the costs resulting from the wear and tear of the medium of exchange. One may note in passing that Marx's discussion of the costs of exchange in terms of the classical notion of unproductive labour provides one of the most immediate links between classical political economy and modern debates on, for example, the relationship between production, transaction costs, and the question whether the latter are best regarded as a cost in the first instance (for example, Barzel 1985; Goldberg 1985).

Conceptual Development

In the 1930s, a number of authors explored the implications of 'selling' or 'marketing' costs as part of the growing literature on advertising, monopolistic competition and the theory of the firm (Braithwaite 1928; Chamberlin 1933; Coase 1937). While the theory of the firm experienced a strong revival three decades later within the emerging new institutional economics (Malmgren 1961; Cheung 1969; Williamson 1970; Alchian and Demsetz 1972), it was Hicks's (1935) explanation of the holding of cash balances on the basis of the costs one incurs when converting assets into cash (Klaes 2000) that provided the strongest immediate impetus to the coining and further differentiation of 'transaction costs', notably via the post-war neo-Keynesian literature, its inventory approach to the transactions demand for money, and the general question of cash balances in general equilibrium theory (Makower and Marschak 1938; Marschak 1950; Baumol 1952; Tobin 1956; Patinkin 1965; Foley 1970; cf. Ulph and Ulph 1975).

The monetary economics literature proceeded largely on the basis of a narrow price-based understanding of transaction costs, although comparative issues that pointed towards broader interpretations became a pressing technical concern in its mature phase (for example, Hahn 1973). Arrow (1965, 1969) provided a crucial conceptual link between the general equilibrium literature on transaction costs and the emerging literature in the new institutional economics by moving from a monetary to a comparative institutional interpretation of transaction costs. By contrast, the economics of property rights and the emerging field of law and economics developed an increasing variety of relational transaction cost interpretations, typically expressed in contractual terms (Coase 1960; Alchian 1965; Demsetz 1967; Calabresi 1968; Posner 1972; Macneil 1981). The overall thrust of the new institutional economics, notably in economic history (Davis and North 1970; North 1985) and through the formulation of a 'transaction cost economics' by Williamson (1975, 1985), has, however, been the spelling out of the details of a comprehensive institutional analysis of institutional arrangements on the basis of comparative institutional interpretations of transaction costs (see also Aoki 2001; Langlois 2002; Greif 2006).

Empirical studies, even on the basis of monetary transaction costs, which one might have expected to be more amenable to direct measurement than broader interpretations, have displayed considerable divergence regarding the level of observed transaction costs. Once one moves beyond narrow monetary transaction cost conceptions, for example to account for observed bid–ask spreads or for violations of the law of one price, one is forced to delimit those dimensions of transaction costs that are not empirically accessible as such and have therefore to be inferred indirectly. In the absence of a commonly agreed empirical definition of transaction costs this will inevitably lead to results that are sensitive

to the particular definition that one employs (Gould and Galai 1974; Fama 1991).

Alternative approaches, largely found in the comparative institutional transaction cost literature, have included attempts to infer macroeconomic transaction costs on the basis of the size of the transaction sector of an economy (Wallis and North 1986), or to derive proxy measures for transaction costs on the basis of transaction characteristics such as asset specificity (Williamson 1991; Nooteboom 1993). The irony of sectoral measures is that economies with less well-developed transaction sectors appear to exhibit lower levels of transaction costs if those costs are measured in terms of sector size, whereas micro-structurally those economies in fact suffer from higher levels of transaction costs due to significant barriers to smooth exchange and coordination of economic activity. Proxy measures, in turn, are at risk of running into difficulty once they seek to embrace institutional transaction cost interpretations, because they have to proceed on the assumption that transactions can be defined in such a way that there is a core to them that remains unaffected across alternative institutional settings. Once one moves to non-market forms of economic coordination, this may become problematic (Masten 1996), although the modularity literature has begun addressing this issue (Langlois 2006).

See Also

- ▶ Adjustment costs
- ▶ Financial intermediation
- ▶ Firm, theory of the
- ▶ Law, economic analysis of
- ▶ New institutional economics
- ▶ Property rights
- ▶ Switching costs
- ▶ Trade costs
- ▶ Vertical integration

Bibliography

Alchian, A.A. 1965. Some economics of property rights. In *Economic forces at work*, ed. A.A. Alchian. Indianapolis: Liberty Press. 1977.

Alchian, A.A. 1991. Development of economic theory and antitrust: A view from the theory of the firm. *Journal of Institutional and Theoretical Economics* 147: 232–234.

Alchian, A.A., and H. Demsetz. 1972. Production, information costs, and economic organization. *American Economic Review* 62: 777–795.

Allen, D.W. 2000. Transaction costs. In *Encyclopedia of law and economics*, vol. 1, ed. B. Bouckaert and G. De Geest. Cheltenham: Edward Elgar.

Anon. 1936. Born of necessity. *Arcadia Tribune*, 12 September, 2. Arcadia: Arcadia Publishing Company.

Aoki, M. 2001. *Toward a comparative institutional analysis*. Cambridge, MA: MIT Press.

Aristotle. 1932. *Politics*. Trans. H. Rackham. London/Cambridge, MA: Heinemann and Harvard University Press, 1967.

Arrow, K.J. 1965. Uncertainty and the welfare economics of medical care: Reply (the implications of transaction costs and adjustment lags). *American Economic Review* 55: 154–158.

Arrow, K.J. 1969. The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation. In *Collected papers*, vol. 2, ed. K.J. Arrow. Oxford: Blackwell. 1983.

Barzel, Y. 1985. Transaction costs: Are they just costs? *Journal of Institutional and Theoretical Economics* 141: 4–16.

Baumol, W.J. 1952. The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics* 66: 545–556.

Braithwaite, D. 1928. The economic effects of advertising. *Economic Journal* 38: 16–37.

Calabresi, G. 1968. Transaction costs, resource allocation and liability rules: A comment. *Journal of Law and Economics* 11: 67–73.

Chamberlin, E.H. 1933. *Theory of monopolistic competition*, 6th ed, 1948. Cambridge, MA: Harvard University Press.

Cheung, S.N. 1969. *The theory of share tenancy*. Chicago: University of Chicago Press.

Clower, R.W. 1995. Axiomatics in economics. *Southern Economic Journal* 62: 307–319.

Coase, R.H. 1937. The nature of the firm. *Economica NS* 4: 386–405.

Coase, R.H. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Coase, R.H. 1972. *Industrial organization: A proposal for research*. Chicago/London: University of Chicago Press.

Coase, R.H. 1974. The choice of the institutional framework: A comment. *Journal of Law and Economics* 17: 493–496.

Coase, R.H. 1988. *The firm, the market, and the law*. Chicago/London: Chicago University Press.

Dahlman, C.J. 1979. The problem of externality. *Journal of Law and Economics* 22: 141–162.

Davidson, M.G. 1896. Friction in economics. In *Dictionary of political economy*, vol. 2, ed. R.H.I. Palgrave. London: Macmillan.

Davis, L.E., and D.C. North. 1970. Institutional change and American economic growth: A first step towards a theory of institutional innovation. *Journal of Economic History* 30: 131–149.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57: 347–359.

Demsetz, H. 1988. *Ownership, control and the firm: The organization of economic activity*. Oxford: Blackwell.

Dixit, A.K. 1996. *The making of economic policy*. Cambridge, MA: MIT Press.

Eggertsson, T. 1990. *Economic behaviour and institutions*. Cambridge: Cambridge University Press.

Fama, E.F. 1991. Efficient capital markets: II. *Journal of Finance* 46: 1575–1617.

Foley, D.K. 1970. Economic equilibrium with costly marketing. *Journal of Economic Theory* 2: 276–291.

Furubotn, E.G., and R. Richter. 2005. *Institutions and economic theory*, 2nd ed. Ann Arbor: University of Michigan Press.

Goldberg, V.P. 1985. Production functions, transactions costs and the new institutionalism. In *Issues in contemporary microeconomics and welfare*, ed. G. Feiwel. Albany: State University of New York Press.

Gould, J.P., and D. Galai. 1974. Transactions costs and the relationship between put and call prices. *Journal of Financial Economics* 1: 105–129.

Greif, A. 2006. *Institutions and the path to the modern economy*. Cambridge: Cambridge University Press.

Grossman, S.J., and O.D. Hart. 1986. The cost and benefit of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94: 691–719.

Hahn, F.H. 1973. On transaction costs, inessential sequence economies and money. *Review of Economic Studies* 40: 449–461.

Hardt, Ł. 2006. Transaction cost economics as a three dimensional externally driven research program. *Studia Ekonomiczne* 48–9: 7–31.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica NS* 2: 1–19.

Holmstrom, B., and P. Milgrom. 1994. The firm as an incentive system. *American Economic Review* 84: 972–991.

Hume, D. 1752. Of money. In *Essays: Moral, political, literary*, ed. E.F. Miller. Indianapolis: Liberty Fund. 1987.

Klaes, M. 2000. The history of the concept of transaction costs: Neglected aspects. *Journal of the History of Economic Thought* 22: 191–216.

Klaes, M. 2001. Begriffsgeschichte: Between the Scylla of conceptual and the Charybdis of institutional history of economics. *Journal of the History of Economic Thought* 23: 153–179.

Klaes, M. 2006. Founding economic concepts. *Storia del Pensiero Economico NS* 3: 23–39.

Kranton, R.E., and D.F. Minehart. 2001. A theory of buyer–seller networks. *American Economic Review* 91: 485–508.

Langlois, R.N. 2002. Modularity in technology and organization. *Journal of Economic Behavior and Organization* 49: 19–37.

Langlois, R.N. 2006. The secret life of mundane transaction costs. *Organization Studies* 27: 1389–1410.

Luttmer, E.G. 1996. Asset pricing in economies with frictions. *Econometrica* 64: 1439–1467.

Macneil, I.R. 1981. Economic analysis of contractual relations: Its shortfalls and the need for a 'rich classificatory apparatus'. *Northwestern University Law Review* 75: 1018–1063.

Makower, H., and J. Marschak. 1938. Assets, prices and monetary theory. *Economica NS* 5: 261–288.

Malmgren, H.B. 1961. Information, expectations and the theory of the firm. *Quarterly Journal of Economics* 75: 399–421.

Marschak, J. 1950. The rationale of the demand for money and of 'money illusion'. *Metroeconomica* 2: 71–100.

Marx, K. 1893. *Das Kapital*, vol. 2. Berlin: Dietz. 1953.

Masten, S.E. 1996. Empirical research in transaction cost economics: Challenges, progress, directions. In *Transaction cost economics and beyond*, ed. J. Groenewegen. Boston: Kluwer.

Ménard, C., and M.M. Shirley. 2005. *Handbook of new institutional economics*. Dordrecht: Springer.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Wien: Braumüller.

Mill, J.S. 1848. *Principles of political economy*, 7th ed, 1909. London: Longmans, Green.

Nooteboom, B. 1993. An analysis of specificity in transaction cost economics. *Organization Studies* 14: 443–451.

North, D.C. 1985. Transaction costs in history. *Journal of European Economic History* 14: 557–576.

North, D.C. 1990. *Institutions, institutional change and economic performance*. Cambridge: Cambridge University Press.

Patinkin, D. 1965. *Money, interest, and prices*, 2nd ed. New York/Tokyo: Harper & Row and Weatherhill.

Posner, R.A. 1972. *Economic analysis of law*. Boston: Little, Brown.

Rutherford, M. 1994. *Institutions in economics: The old and the new institutionalism*. Cambridge: Cambridge University Press.

Say, J.B. 1803. *A treatise on political economy*, 4th ed. trans. C. R. Prinsep, Philadelphia: Lippincott, Grambo & Co, 1855.

Scitovsky, T. 1940. A study of interest and capital. *Economica NS* 7: 293–317.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.

Ulph, A.M., and D.T. Ulph. 1975. Transaction costs in general equilibrium theory: A survey. *Economica* 42: 355–372.

Wallis, J.J., and D.C. North. 1986. Measuring the transaction sector in the American economy, 1870–1970. In *Long term factors in American economic growth*, ed. S.L. Engerman and R.E. Gallman. Chicago/London: University of Chicago Press.

Williamson, O.E. 1970. *Corporate control and business behavior*. Englewood Cliffs: Prentice-Hall.
Williamson, O.E. 1975. *Markets and hierarchies*. New York: Free Press.
Williamson, O.E. 1985. *The economic institutions of capitalism*. New York: Free Press.
Williamson, O.E. 1991. Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly* 36: 269–296.
Williamson, O.E. 2000. The new institutional economics: Taking stock, looking ahead. *Journal of Economic Literature* 38: 595–613.

Transfer of Technology

Wolfgang Keller

Abstract

This article discusses the transfer of technological knowledge with a focus on firms in different countries. The dominant form of such technology transfer is not market transactions at arm's length but technological learning externalities. The article reviews the evidence for technological learning from importing, exporting, and foreign direct investment activities.

Keywords

Asymmetric information; Foreign direct investment; Foreign direct investment spillovers; Growth and international trade; Innovation; Learning externalities; Productivity growth; Research and development; Technology spillovers; Trade costs; Transfer of technology

JEL Classifications

O3

Asymmetric information often makes the buying, selling or licensing of technology unfeasible. Instead, international technology transfer tends to occur through non-market channels. In the United States, for example, the figures on trade in services in the balance of payments are much smaller than estimates on US technological externalities (McNeil and Fraumeni 2005). These externalities are called technology spillovers.

A good starting point is the model of international technology transfer by Howitt (2000). There are many intermediate good sectors each characterized by its own level of technology. Different countries and sectors employ different technologies based on domestic innovation and technology transfer from abroad. At any given time t, the technology frontier, A_t^{max}, denotes the highest technology level across all countries and sectors. The technology frontier is growing because innovations worldwide push it out over time. An innovation in a particular sector brings this sector's productivity up to the technology frontier. This means that the model includes international and inter-sectoral technology spillovers, since if there had been many innovations in other countries or sectors, the jump to the technology frontier for sector $i$is larger than if there had been few innovations elsewhere.

With productivity in successfully innovating sectors jumping to the frontier level, a country's average productivity across sectors (denoted by A_t) rises. In Howitt's model, average productivity changes according to

$$\dot{A}_t = \lambda r_t \left(A_t^{\max} - A_t\right), \tag{1}$$

where r_t is a measure of domestic R&D investment, and $\lambda > 0$ is a parameter. In (1), the change in A_t is positively related to domestic R&D, and it is positively related to the average technology gap $\left(A_t^{\max} - A_t\right)$. Equation (1) leads to a common long-run growth rate shared by all countries; however, those investing more in R&D will enjoy relatively high productivity levels, all else being equal.

The literature emphasizes that technology transfer is facilitated by firms' international activity, though to date this has not been comprehensively articulated at the theoretical level. International trade and foreign direct investment have long been discussed as some of the most important channels, and the econometric evidence is discussed in the following (see also Keller 2004).

First, imports may lead to technology transfer. Coe and Helpman (1995) test the prediction of the trade and growth models of Grossman and Helpman (1991) and Rivera-Batiz and Romer (1991), in which foreign R&D creates new intermediate inputs and perhaps generates spillovers for the home country through importing activity. Output is produced with labour and differentiated capital goods that enter in a constant elasticity of substitution (CES) function. The intermediate product range in each country is expanded through R&D, and countries can benefit from other countries' R&D by importing foreign-developed intermediate goods. Under certain assumptions, a country's productivity f is given by

$$\ln f = \ln B + \alpha \ln n^e, \qquad (2)$$

where n^e is the range of intermediate goods employed in the country, and $B > 0$ is a parameter. According to (2), productivity is positively related to the intermediate products range employed in this country.

A country's demand for intermediates will depend on bilateral trade barriers and transport costs. Coe and Helpman (1995) distinguish between foreign and domestic products, since trade costs are often very different between these

$$\mathrm{In} f_c = \alpha_c + \beta^d \mathrm{In}\, n_c + \beta^f \mathrm{In}\, n_c^f + \varepsilon_c, \qquad (3)$$

where n_c^f is defined as the bilateral import share weighted R&D of country c's trade partners: $n_c^f = \sum_{c' \neq c} m_{cc'} r_{c'}$. This captures the prediction that if a country imports primarily from high-R&D countries, it is likely to benefit more from foreign technology than if it imports primarily from low-R&D countries.

While Coe and Helpman (1995) estimate a positive and quantitatively large effect from import-weighted foreign R&D, subsequent work by Keller (1998) shows that this per se cannot be taken as evidence for imports-related international technology transfer. Instead of using data on actual bilateral imports, Keller (1998) conducts robustness checks with two alternative foreign variables based on randomly created shares as well as no shares at all: $\tilde{n}_c^f = \sum_{c' \neq c} \mu_{cc'} r_{c'}$ *and* $\tilde{\tilde{n}}_c^f = \sum_{c' \neq c} r_{c'}$. Since these alternative variables yield similar or even stronger results, as in Coe and Helpman (1995), the observed import patterns cannot explain the estimated effects.

By making progress on a number of fronts, further work has produced robust evidence for imports-related technology transfer. First, Xu and Wang (1999) and Keller (2000) note that as a matter of theory, foreign technology spillovers are the result of capital goods trade, and not aggregate trade, which Coe and Helpman (1995) use to construct their import shares. Xu and Wang (1999) show that if the foreign variable n_c^f is based on bilateral capital goods trade shares, it performs better than both Coe and Helpman's (1995) original and Keller's (1998) alternative variables. Moreover, since most R&D is conducted in a relatively small part of manufacturing, imports-related technology transfer effects are relatively difficult to estimate with country-level data, as in Coe and Helpman; in a study among Organisation for Economic Co-operation and Development (OECD) countries at the two- and three-digit industry level, Keller (2002) finds robust evidence that imports are a channel for international technology transfer.

Second, the major question regarding exports is whether firms learn about foreign technology through exporting experience. There is abundant evidence that in a given cross-section exporters are on average more productive than non-exporters (Bernard and Jensen 1999). This does not address the question whether exporting firms become more productive because of learning effects associated with exporting, or whether firms that are more productive to begin with export more. According to much anecdotal evidence, firms do benefit from interacting with the foreign customer, for instance because the latter imposes higher product quality standards than the domestic customer, while at the same time providing information on how to meet the higher standards. The econometric evidence is more mixed, however.

While learning-by-exporting has been emphasized primarily for low- and middle-income countries' firms, there is in principle no reason why it is limited to these countries. Bernard and Jensen (1999) analyse learning-by-exporting using data on US firms. In studying the performance of four

different sets of firms – exporters, non-exporters, starters and quitters – separately, Bernard and Jensen (1999) do not model export market participation explicitly. They find that labour productivity growth is about 0.8 per cent higher among exporters than non-exporters. This estimate is fairly small, and it becomes even smaller (and insignificant) for longer time horizons. At the same time, this is conditional on plant survival. Bernard and Jensen also show that exporters are ten per cent more likely to survive than non-exporters. This difference is indicative of higher productivity growth for exporters than non-exporters, because low productivity growth is the primary reason for plants failing. Thus, there may be learning-from-exporting effects that amount to more than 0.8 per cent, although it is not clear whether they are substantial.

The paper by Clerides et al. (1998) provides evidence on learning externalities from exporting using micro data from Columbia, Morocco and Mexico. By estimating simultaneously a dynamic discrete choice equation that determines export market participation, these authors take into account the consideration that it is on average the already-productive firms that self-select into the export market. The export market participation decision is given by

$$y_{it} = \begin{cases} 1 \quad if \quad 0 \le \beta^x X_{it} + \beta^e e_{it} + \sum_{j=1}^{J} \beta_j^c \ln\left(AVC_{it-j}\right) \\ \qquad\qquad + \sum_{j=1}^{J} \left(F^0 - F^j\right) y_{it-j} + \eta_{it} \\ 0 \quad otherwise, \end{cases} \tag{4}$$

and any learning from exporting effects are uncovered by simultaneously estimating an autoregressive cost function

$$\begin{aligned} \ln(AVC_{it}) = \gamma_0 &+ \sum_{j=1}^{J} \gamma_j^k \ln(K_{it-1}) \\ &+ \gamma^e \ln(e_t) + \sum_{j=1}^{J} \gamma_j^c \ln\left(AVC_{it-j}\right) \\ &+ \sum_{j=1}^{J} \gamma_j^y y_{it-j} + v_{it} \end{aligned} \tag{5}$$

In Eq. (4), y_{it} is the export indicator of plant i in period t, X_{it} is a vector of exogenous plant characteristics, e_t is the exchange rate, AVC_{it} are average costs, K_{it} is capital, and $F^0$and F^j are sunk costs of export market participation.

Equation (4) states that one only sees a plant exporting if the profits from doing so are greater than from not exporting (the latent threshold is expressed in terms of observables). Equation (5) asks whether past exporting experience reduces current cost (captured by the parametersγ_j^y), conditional on past costs and size (proxied by capital). Clerides et al. (1998) show results for the three countries separately, and also by major industry, using maximum likelihood (MLE) and generalized method of moments (GMM) methods. These estimations show no significant positive effects from past exporting experience on current cost. These authors' descriptive plots of average cost before and after export market entry support this conclusion. Thus, exporting does not facilitate technology transfer. According to the Clerides et al. (1998) analysis, exporters are more productive, but that is because they self-select themselves into the export market.

Using similar methods, van Biesebroeck (2005) has revisited the issue by studying productivity dynamics of firms in nine African countries. In contrast to Clerides, Lach and Tybout, he estimates that the firm starting to export boosts productivity by about 25 per cent on average in his sample. Van Biesebroeck (2005) also estimates that the higher productivity growth of exporters versus non-exporters is sustained. By employing instrumental-variable and semi-parametric techniques as alternative ways to deal with the selection issue, Van Biesebroeck's analysis is more comprehensive than most. His analysis generally supports the notion that exporting leads to the transfer of technological knowledge. In trying to reconcile his findings with some of the earlier results, Van Biesebroeck shows that part of the difference in productivity growth between exporters and non-exporters appears to be due to unexploited scale economies for the latter. This suggests that at least in part his results are due to constraints imposed by demand, and not due to technology transfer in the sense of an outward

shift of the production possibility frontier at all levels of production. We need richer data to make further progress on distinguishing these hypotheses.

Third, foreign direct investment (FDI) has long been considered as an important channel for technology transfer. Among the possible mechanisms are knowledge spillovers, labour turnover, linkages, and advanced specialized inputs. Also, multinational companies are well known to be more productive and do more R&D than purely domestic firms, so they are likely sources for such productivity benefits. Moreover, governments all over the world spend large amounts of resources to attract affiliates of multinationals to their jurisdiction. If this is rational economic policy, there ought to be large technology transfers associated with FDI.

Numerous studies have estimated FDI spillovers since 1970. Recently, authors focus on panel data analysis with micro data, since this reduces problems resulting from unobserved heterogeneity across firms and sectors. Typically, a general relationship between productivity growth of domestically owned firms (Δf) and a measure of the change in inward FDI (ΔFI) is specified in order to uncover evidence for FDI spillovers:

$$\Delta f_{ist} = \beta X' + \gamma \Delta FI_{ist} + u_{ist}, \qquad (6)$$

Here, X is a vector of control variables, u is a regression error, and i, s and t are firm, industry and time subscripts, respectively. The spillover parameter γ is estimated positive if productivity growth of firms in industries that have experienced large increases in FDI exceeds that of firms in industries where FDI has grown little.

Until about 2002, many authors concluded, by and large, that there is no evidence for FDI spillovers. This is also reflected in a number of surveys (Lipsey and Sjöholm 2005; Görg and Greenaway 2004; Hanson 2001). The paper by Aitken and Harrison (1999) even estimates a *negative* relationship between FDI and productivity for a sample of Venezuelan plants. Since technology learning spillovers can hardly be negative, the analysis probably picks up something else. One possibility, first suggested by Aitken and Harrison (1999), is that the negative coefficient is due to increased competition for local plants through foreign entry. Alternatively, it could be due to endogeneity, if FDI flows to sectors in which firms are relatively weak.

The paucity of evidence has led some to look elsewhere: if there are no spillovers to domestic firms in the same industry, perhaps they exist for domestic suppliers of multinational firms? Contractual relations between foreign-owned affiliates and their domestic suppliers suggest that the technology transfer could, in principle, be specified and paid for – in which case these are not externalities. However, there could be learning effects on top of this. The paper by Smarzynska Javorcik (2004) finds evidence consistent with vertical spillovers between firms in different industries in Lithuania, but no within-industry spillovers.

Haskel et al. (2002) and Keller and Yeaple (2003) have returned to the original question of spillovers from FDI in a given industry. The former estimate an equation like (6) for FDI into the United Kingdom, and the latter for FDI into the United States. Haskel et al. (2002) estimate positive spillovers, which, however, are relatively small, as the authors note. More importantly, these authors, as is the case for Smarzynska Javorcik (2004), do not fully address the possibility that FDI inflows may be endogenous.

The first paper to show that multinationals can cause economically large productivity benefits to domestically owned firms is Keller and Yeaple (2003). These authors deal with endogeneity concerns using instrumental variable techniques, and they employ the by now well-known Olley and Pakes (1996) method of computing firm productivity. The resulting estimates imply an influence of FDI on productivity growth that is much larger than in existing studies. Using data on about 1300 US manufacturing firms for the years 1987–96, they estimate that FDI spillovers explain about 11 per cent of productivity growth during this time.

Keller and Yeaple (2003) also reconcile their results with earlier studies that have found no evidence for FDI spillovers. For one, FDI spillovers are heterogeneous, with much stronger effects in the relatively high-technology industries. Secondly, large FDI spillovers are estimated

only with the high-quality data on FDI by industry they employ. If, instead, Keller and Yeaple (2003) use FDI data similar to that more commonly available in other studies, they too estimate only a small or zero effect.

Thus, in contrast to the earlier literature, the most recent micro productivity studies tend to estimate positive, and in some cases also economically large spillovers associated with FDI. As one would expect, the effects are heterogeneous across industries, with stronger effects in relatively high-tech industries. It is not clear yet whether strong FDI spillovers occur only in relatively rich but not in relatively poor countries. Another of Keller and Yeaple's findings, that relatively weak firms benefit more from FDI than stronger firms, suggests that FDI spillovers are not limited to rich countries, where firm productivity tends to be relatively high.

To conclude, recent experience with research on the channels of technology transfer in all three areas, imports, exports and FDI, clearly shows that it is crucial to have access to detailed information or at least proxy variables on the technology being transferred. Given that much of technology transfer is associated with externalities, this may be the single most important issue that future work needs to address.

See Also

- ▶ Countertrade
- ▶ Diffusion of Technology
- ▶ Foreign Direct Investment
- ▶ Trade, Technology Diffusion and Growth

Bibliography

Aitken, B., and A. Harrison. 1999. Do domestic firms benefit from FDI? Evidence from Venezuela. *American Economic Review* 89: 605–618.

Bernard, A., and B. Jensen. 1999. Exceptional exporter performance: Cause, effect, or both? *Journal of International Economics* 47: 1–26.

Clerides, S., S. Lach, and J. Tybout. 1998. Is learning by exporting important? Micro-dynamic evidence from Columbia, Mexico, and Morocco. *Quarterly Journal of Economics* 113: 903–947.

Coe, D., and E. Helpman. 1995. International R&D spillovers. *European Economic Review* 39: 859–887.

Görg, H., and D. Greenaway. 2004. Much ado about nothing? Do domestic firms really benefit from FDI? *World Bank Research Observer* 19: 171–197.

Grossman, G., and E. Helpman. 1991. *Innovation and growth in the global economy*. Cambridge, MA: MIT Press.

Hanson, G. 2001. *Should countries promote FDI? G-24 discussion paper series*. New York/Geneva: United Nations.

Haskel, J., S. Pereira, and M. Slaughter. 2002. *Does inward FDI boost the productivity of domestic firms?* Working Paper No. 8724. Cambridge, MA: NBER.

Howitt, P. 2000. Endogenous growth and cross-country income differences. *American Economic Review* 90: 829–844.

Keller, W. 1998. Are international R&D spillovers trade-related? Analyzing spillovers mong randomly matched trade partners. *European Economic Review* 42: 1469–1481.

Keller, W. 2000. Do trade patterns and technology flows affect productivity growth? *World Bank Economic Review* 14: 17–47.

Keller, W. 2002. Trade and the transmission of technology. *Journal of Economic Growth* 7: 5–24.

Keller, W. 2004. International technology diffusion. *Journal of Economic Literature* 42: 752–782.

Keller, W., and S. Yeaple. 2003. *Multinational enterprises, international trade, and productivity growth: Firm-level evidence from the United States*. Working Paper No. 9504. Cambridge, MA: NBER.

Lipsey, R., and F. Sjöholm. 2005. The impact of inward FDI on host countries: Why so different answers? In *Does Foreign direct investment promote development?* ed. T.H. Moran, E.M. Graham, and M. Blomström. Washington, DC: Institute for International Economics.

McNeil, L., and B. Fraumeni. 2005. *International trade and economic growth: A possible methodology for estimating cross-border R&D spillovers*. BEA Working Paper 2005–03. Washington, DC.

Olley, S., and A. Pakes. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64: 1263–1297.

Rivera-Batiz, L., and P. Romer. 1991. Economic integration and endogenous growth. *Quarterly Journal of Economics* 106: 531–555.

Smarzynska Javorcik, B. 2004. Does FDI increase the productivity of domestic firms? In search of spillovers through backward linkages. *American Economic Review* 94: 605–627.

Van Biesebroeck, J. 2005. Exporting raises productivity in Sub-Saharan African manufacturing firms. *Journal of International Economics* 67: 373–391.

Xu, B., and J. Wang. 1999. Capital goods trade and R&D spillovers in the OECD. *Canadian Journal of Economics* 32: 1258–1274.

Transfer Payments

Robert J. Lampman

A transfer transaction is unlike an exchange transaction. The latter, which is the main concern of market economists, involves two trading partners both of whom give up something of value in search of mutual gain. The former involves a donor and a recipient, with the donor giving up something of value without receiving anything in return. Transfers can be made between one person and another or from one organization such as a government to another. The transaction can be explicit as in the case of a stated gift to a certain person or quite diffuse as in the case of a subsidy to anyone who produces or consumes a specific consumer good. Transfers, which may take the form of income or wealth, can be voluntary or involuntary and may be motivated either by altruism of the donor of malevolence of the recipient.

The study of transfer payments can take one in many directions. However, the most common direction is that given by national income accountants in reference to transfer payments to persons. In the United States this term is restricted mainly to payments from government and business to the personal or household sector (United States 1981, p. xi). Government payments recorded under this heading include cash benefits paid out under social insurance and public assistance programmes as well as programmes for government employees, veterans and students. Also included are certain cash payments to nonprofit organizations which are considered part of the household sector. In addition to those cash payments, some in-kind benefits, notably health care under social insurance and food stamps, are included. The total of government transfer payments to persons equalled about 10 per cent of GNP in the 1980s.

A much smaller amount is recorded as business transfers to persons, which include corporate gifts to non-profit organizations, consumer bad debts, liability claims for personal injuries, thefts and forgeries, and cash prizes. The national income accounts also show a similarly small flow of transfer payments to foreigners from the domestic household sector and the government sector.

United States national income accounting practice emphasizes monetary benefits, it uses a one-year accounting period to help distinguish a transfer from an exchange, and it excludes transfers which do not cross sectors but occur within the polyglot household sector. This accounting practice is criticized both for items it leaves out and for some of those it includes. Questions raised include the following. Should social insurance benefits paid for by the recipient in a prior year be classified as deferred wages rather than as transfers? Why are health care under social insurance and food stamps included, while many other non-monetary benefits – for example, education – are not? Why are benefits paid out by private pension funds and group health insurance plans not counted as transfer payments to persons? Should corporations' payments of interest and dividends be classified as business transfer to persons on the grounds that unlike the return to physical capital owned by the corporation, the payment by the corporation to the owner of financial assets may not relate to current changes in real capital and land supplied to the productive process? (This classification is followed in the United Nations' System of National Accounts, United Nations 1968.) Should capital gains and losses be recognized as transfers? (Eisner 1984).

These questions are all difficult ones for national income accountants, who see their role primarily as one of measuring output produced in a year. For this purpose the activities of business and government are of major importance and the flow of income to and within the household sector is of lesser significance. To the extent that it is desirable to show more of the income received by persons, modifications of the personal or household sector and a broader definition of transfer are needed. Such changes include (1) dividing this sector into separate subsectors for consumer units, nonprofit institutions, and private pension and group insurance organizations, and (2) identifying more in-kind government benefits as transfer to persons (Ruggles and Ruggles 1982).

While national income accountants have been slow to move on this question, other scholars have

take steps to widen the inquiry. One group of economists has undertaken to show how all taxes and all government spending, including that for public goods, affect the distribution of income. Public goods are, of course, particularly difficult to allocate among specific families. This kind of study relies upon tax records, records of expenditures under public programmes, and household surveys of family income and expenditures classified by size and composition of families (Gillespie 1965; Reynolds and Smolensky 1977). In general, such studies find that while tax systems are only slightly redistributive from rich to poor, government expenditures are more clearly so.

Still another group of economists, this one led by Boulding (1973), see the phenomenon of transfer as much more fundamental than is suggested by any list of formal public and private redistributive programmes. They envision every economy as comprising two co-existing systems, a market system and a transfer or 'grants' system. The latter enters into almost all producer as well as consumer decisions. Grants arise not only from government expenditures but also, implicitly, from such regulatory actions as restrictions on trade, occupational licensing, and the setting of wage rates and work rules by collective bargaining.

A third group can be identified as staking out a middle ground for the study of transfers. This middle ground is based on a concept of a redistributive budget or a social budget. It starts with a list of explicit, as distinct from Boulding's implicit, transfers, and it excludes transfers of public goods. The private goods that go to specified consumer units include cash benefits and such in-kind benefits as education and health care. A number of nations, including the United States and the Federal Republic of Germany, have made use of social budgets, and the OECD recently developed a standard reporting pattern for use in comparing the social spending of its member nations (OECD 1985). Social spending grew faster than total government spending in all OECD nations from 1960 to about 1975. The ratio of social spending to GNP averaged 25.6 per cent in those nations in 1981, with a low of 13.4 per cent in Greece and a high of 37.6 per cent in Belgium. The United States and the United Kingdom were each a few percentage points below the average.

Lampman (1984) has proposed a two-way expansion of the social budget to make it more useful for comparison across nations and over time. First, he adds the role played by private alternatives to government delivery of social benefits. In so doing he implements the subsectoring of the household sector referred to above in the discussion of national income accounting. He conceives of transfers as flowing from families to families as direct gifts from one family to another and also via intermediary organizations including government, private philanthropies such as churches and private schools, and private pension funds and group health insurance plans. Secondly, he adds information to show the other side of the budget, detailing which families pay for the total of current benefits received.

Using this scheme of an expanded social budget, Lampman finds that the ratio of social spending to GNP in the United States in 1978 was 27.6 per cent inclusive of 19.6 per cent via government intermediaries. The private role of 8 per cent was made up of 4 per cent interfamily direct giving, 3.6 per cent private group insurance, and 0.6 per cent philanthropic giving. Over half of all social benefits as defined take the form of cash, and most of the in-kind benefits are education and health care services. The largest share of cash benefits goes to the retired aged and about a third of all benefits go to the pre-transfer poor.

It is important to relate the definition of transfers from and to persons to the widespread recognition that they may influence patterns of consumption, saving, investment in human capital, family composition, and work effort. Improved quantitative information may help to resolve questions of why social spending has increased. Is it due to market failure to deal with externalities and high costs of information and transactions, or is it due to government failure to identify the optimal quantities of transfer?. Further, such information may assist in the illumination of the social benefits and social costs that flow from recent expansion of social spending (Wilson and Wilson 1982).

See Also

- ▶ Negative Income Tax
- ▶ Poor Law, old
- ▶ Social Security

Bibliography

Boulding, K.E. 1973. *The economy of love and fear: A preface to grants economics*. Belmont: Wadsworth.

Eisner, R. 1984. Transfers in a total income system of accounts. In *Economic transfers in the United States*, Studies in income and wealth, vol. 49, ed. M. Moon, 9–35. Chicago/London: University of Chicago Press.

Gillespie, W.I. 1965. The effects of public expenditures on the distribution of income. In *Essays in fiscal federalism*, ed. R.A. Musgrave. Washington, DC: Brookings Institution.

Lampman, R.J. 1984. *Social welfare spending: Accounting for changes from 1950 to 1978*. Orlando: Academic.

OECD. 1985. *Social expenditure: 1960–1990*. Paris: OECD.

Reynolds, M., and E. Smolensky. 1977. *Public expenditures, taxes, and the distribution of income: The U.S. 1950, 1961, 1970*. New York: Academic.

Ruggles, R., and N.D. Ruggles. 1982. Integrated economic accounts for the United States, 1947–1980. *US Department of Commerce, Survey of Current Business* 62: 1–53.

United Nations, Department of Economic and Social Affairs. 1968. System of national accounts. In *Studies in methods*, Series F, No. 2, rev. 3. New York.

US Department of Commerce. 1981. *The National Income and Product Accounts of the United States 1929–76. Statistical tables*. Washington, DC: Government Printing Office.

Wilson, T., and D.J. Wilson. 1982. *The political economy of the welfare state*. London: Allen & Unwin.

Transfer Pricing

Jack M. Mintz

Abstract

Governments establish tax rules for setting transfer prices for non-arm's length transactions made by multinationals, following guidelines established by the OECD (1995) under Article 9 of the OECD Model Tax Convention. Various methodologies have been established; the first preference is determining comparable uncontrolled prices according to the arm's length principle. Given the difficulties of achieving comparable transactions in determining price, margins or profitability, other methods, such as allocating profits among members of a corporate group according to a formula, have instead been relied upon for multi-jurisdictional corporate income taxation in some circumstances.

Keywords

Arm's length prices; Capital intensity; Comparable uncontrolled price; Competent authority; Cost-plus margin; Double taxation; Formulary apportionment; Resale-minus margin; Tax treaties; Transactional net margin method; Transfer pricing

JEL Classifications

H2

Transfer prices are established for goods and services sold between related parties among members of a multinational group. With the growth of cross-border transactions by multinational businesses, tax authorities increasingly deal with issues related to the proper assessment of transfer prices to measure corporate income, since multinationals may use transfer prices to reduce worldwide payment of taxes by shifting income from high-tax to low-tax jurisdictions.

For example, by charging a high price for goods sold to an affiliate purchaser in a high-tax country, the purchaser's reported profits are reduced while higher profits are reported by the vendor affiliate in the low-tax country. The taxes saved by reporting higher costs in the high-tax country are in excess of the additional tax paid on extra income earned in the low-tax country. Similarly, reporting a lower price on goods sold by a company operating in a high-tax jurisdiction to related parties' purchasers located in low-tax jurisdictions also reduces worldwide taxes.

Legal Framework of Transfer Pricing

Governments enact legislation to assess profits earned by multinational companies operating in their jurisdiction based on the 'arm's length principle', which is a price that would be charged by unrelated businesses undertaking a transaction under similar facts and circumstances. As a basis for national legislation, assessment and adjudication, legislation may draw from the guidelines established by the Organisation for Economic Co-operation (1995) for transfer pricing. Each government is ultimately responsible for the development of its tax policy and process, and many require companies to document contemporaneously their pricing methodologies when determining their taxable profits in a jurisdiction.

When a multinational is reassessed by a country for its transfer prices, it could be faced with double taxation of income if other countries choose a different transfer pricing methodology or disagree on the quantum of the inter-company transaction. Many governments have entered into tax treaties that allow corporations to seek an alternative process for double taxation relief (competent authority). In many countries, when taxpayers enter the competent authority programme for double taxation relief they may be obliged to give up their right of appeal and access to tax courts through domestic channels.

Economic Value of Transfer Prices

Even without taxation, transfer prices would be set by multinationals with the objective to improve business performance by efficiently allocating resources amongst the various competing segments of a firm's value chain, and to appropriately reward employees and shareholders of the multinational group. The pricing strategy could therefore distort resource allocation decisions and profits along with determining managerial compensation or the amount of income earned by minority shareholders owning shares in the parent company or affiliates. To help align the interests of the managers with shareholders to maximize shareholder value and to protect the interests of minority shareholders, transfer prices may be set to reflect market conditions, including a reasonable approximation of the arm's length price that would be established between two unrelated parties. Some multinationals have kept two sets of books – one for accounting and another for tax purposes – so that control can be separated from tax motives for establishing transfer prices.

Determining the appropriate price for transfer prices is not an easy task, since prices of comparable transactions can be difficult to obtain. Many transactions involve intangible assets and intellectual property related to research and development, marketing and trademarks, which may be unique to a specific firm and embedded and intertwined with service and tangible goods transfers. Risks also need to be compensated through pricing. Costs are influenced by the size of the transaction and market conditions, including the degree of competitiveness, and regulations and consumer preferences and factor endowments also affect pricing. Finding comparable prices is a difficult matter at best, so it is not uncommon for tax authorities to challenge the transfer price methodology used by multinational companies in assessing taxable income earned in a jurisdiction.

Methods for Determining Transfer Prices

Countries have generally agreed upon a ranking of methodologies to determine transfer prices using the OECD guidelines (1995) although the United States has adopted a 'best method' approach that allows taxpayers to select the most appropriate method. The OECD guidelines identified three transactional and two profit-based methods. The transactional methods include the comparable uncontrolled price method (CUP), the resale-minus price method (RPM), and the cost-plus method (CPLM). The two profit based methods are the profit splits method (PSM) and the transactional-net-margin method (TNMM). The CUP method as observed on transactions between unrelated parties is viewed as the most appropriate methodology for measuring transfer prices for non-arm's length transactions made among members of the multinational group. However, to

employ a CUP, corporations must show that the transaction that they have chosen has similar terms and conditions to those of a related-party transaction including quality and reliability, availability of volume of supply, provision of services, licensing, type and characteristics of property (patent, trademark or know-how), functions undertaken by the enterprise, business strategies, risks and market conditions.

It may be quite difficult to find CUPs that can be adjusted to reflect all potential differences. Therefore, other methodologies might be used instead.

Two other transactional-based methods include the resale-minus and cost-plus approach. The resale-minus method is typically used for distribution companies in that an inter-company price is determined by subtracting a gross profit margin from the product's resale price. An alternative method is the cost-plus method, which would be appropriate for a manufacturer to consider. A profit margin would be added to manufacturing and other costs to determine the cost-plus margin on goods and services sold from one affiliate to another in the multinational group.

With both the resale-minus and cost-plus approach, the multinational's margin on related party sales or costs can be compared with similar transactions made between unrelated parties. However, as with the CUP methodology, care must be taken to put transactions on a comparable basis. For example, companies might have different capital–labour ratios in producing goods and services. Therefore, it would not be surprising if resale-minus or cost-plus margins, reflecting the ratio of profits to costs, tend to be higher for companies relying on capital-intensive techniques of production. For comparability, an adjustment for capital intensity would be needed.

Given that corporate income is a payment made to shareholders after the deduction of borrowing costs, profit-based methods are used as an approach when it is not appropriate to use transactional approaches. The split-profit method would result in a transfer price for a transactions where both parties to a transaction own valuable and significant intangible assets. The transactional net margin method (TNMM) determines the price of an inter-company transfer of goods and services by comparing an associate affiliate's operating profits, in relation to an appropriate base (sales, costs or assets for example), to profits earned by uncontrolled firms. To establish comparability, adjustments are needed for differences in the age of capital (in part due to the distorting impact of inflation on profits) and risk.

Economic Aspects of Transfer Pricing

When businesses choose transfer prices that vary from the 'true' price, they trade off the benefits of tax reduction with non-tax costs, including distorting managerial behaviour or greater expected cost of reassessment by tax authorities (Haufler 2001). Thus, the greater the tax savings from shifting income from high to low transactions, the more the transfer prices will be distorted for tax purposes. Governments may counteract transfer pricing by cutting corporate income tax rates or pursuing more aggressively transfer pricing practices of multinationals through the legal process.

With tighter transfer pricing, companies find that other income-shifting approaches for reducing worldwide taxes can be simpler, such as shifting debt, leasing, insurance, licensing and other deductible expenses to high-tax countries with income report by affiliates operating in tax havens. Transfer pricing litigation may result if the interest rates, fees and royalties charged are not justified at market rates.

Allocation or Apportionment Methods as a Substitute

Given the constraints in assessing comparable prices, margins and profits, tax authorities will sometimes rely on other approaches to assessing corporation taxes rather than assess transfer prices to determine accounting income earned in a country. One approach is to assess a share of the worldwide income earned by a multinational group that would be allocated or apportioned to a specific jurisdiction (formulary apportionment). The share of profits apportioned to a jurisdiction could be

based on one or several factors, including payroll, capital and sales as a portion of worldwide amounts. The allocation or apportionment method (Martens-Weiner 2006) is used in some federal states, including Canada, Switzerland and the United States, to avoid the necessity of determining accounting profits using the transaction approach under separate accounting. It is also used by California to assess its corporate income tax on multinationals operating in the state (Californian income is assessed by multiplying the tax rate by income, which is a percentage of worldwide income; the percentage is determined by a formula). Some experts have advocated the use of the apportionment method at the international level. At the present time, in 2007, member states of the European Union have been debating proposals to consolidate corporate income tax bases with an apportionment method that would allocate profits to each member state.

See Also

- ▶ Tax Havens
- ▶ Tax Treaties

Bibliography

Haufler, A. 2001. *Taxation in a global economy*. Cambridge: Cambridge University Press.

Martens-Weiner, J. 2006. *Company tax reform in the European Union*. New York: Springer.

OECD (Organisation for Economic Co-operation and Development). 1995. *Transfer pricing guidelines for multinational enterprises and tax administrations*. Paris: OECD.

Transfer Problem

Philip Brock

Abstract

A financial transfer of wealth between countries necessitates adjustments in expenditure, production, and relative prices that collectively comprise the transfer problem. Since the 1920s the transfer problem arising from war reparations and other unrequited transfers has occupied the attention of Keynes, several Nobel laureates, and other distinguished economists. The early literature centred on static two-country models of transfers, while more recent research has highlighted the intertemporal dimension of the transfer problem.

Keywords

Infinite horizons; Intertemporal models; Intertemporal terms of trade; Overlapping generations model; Reparations; Representative agent; Terms of trade; Transfer paradox; Transfer problem; Walras's Law

JEL Classifications

F1

Development of the theoretical literature on the transfer problem mirrors the historical context within which the issue first arose. In 1919, as part of the Treaty of Versailles that followed the First World War, Germany was required to make reparations payments to the European powers to which it surrendered (see Eichengreen 1986). Initially, much discussion of Germany's capacity to pay proceeded on the basis of constant international prices and assumed that governments could automatically engineer the required changes in spending on traded goods at home and abroad. But Keynes (1929), in an article which introduced the phrase 'transfer problem' into the professional literature, argued that a country required to make a fixed transfer of purchasing power to another would suffer a secondary burden in the form of a further decline in its purchasing power due to an induced deterioration in its international terms of trade. Ohlin (1929) argued in response that a secondary benefit – or terms of trade improvement – was as likely to occur due to expenditure effects and the presence of non-traded goods (see especially Mundell 2002).

Ohlin's central insight can be stated simply, following Pigou (1932), using a two-country,

two-commodity model of international trade. To highlight the central point, assume initially that production of both goods is exogenously given and that all income is devoted to consumption. If the markets for both commodities clear, then by Walras's Law we need only consider one, for example the good exported by the home country, where that good is denoted x. The total supply of x must equal the sum of domestic and foreign demands:

$$S + S^* = D(y, p) + D^*(y^*, p)$$

where S and S^* designate domestic and foreign supplies, taken as exogenous for the moment (with asterisks denoting foreign values throughout), and D and D^* domestic and foreign demands (each of which depends on real income and relative prices). Now assume that an amount T of purchasing power is transferred from the home to the foreign country. Domestic demand for x falls by $D_y T$ where D_y is the home country's marginal propensity to consume x out of income, while foreign demand for this good rises by $D^*_{y*} T$ where D^*_{y*} is the foreign country's marginal propensity to consume x, also its marginal propensity to import. Equilibrium in the market for x at the initial prices requires that an export surplus in the amount T results from income effects alone – in other words, that $\left(D_y + D^*_{y*}\right) T = T$, or $D_y + D^*_{y*} = 1$. If $D_y + D^*_{y*} < 1$, the combined marginal propensities to spend on x out of income are too small to generate a sufficient surplus at initial relative prices. This is the 'orthodox' case in which the transfer-making country's terms of trade deteriorate, creating a secondary burden (Samuelson 1952, 1971). If $D_y + D^*_{y*} > 1$, then to the contrary the transfer-making country's terms of trade improve.

Once adjustments in production are introduced, the terms of trade will deteriorate when the bias in tastes in each country towards consumption of the exportable good is greater than the bias in production due to international differences in factor endowments or technologies (Jones 1975). Transport costs, by increasing the correlation within countries of patterns of production and consumption, reinforce the orthodox presumption of a secondary burden (Samuelson 1952). Introducing a third country adds an additional set of supply and demand elasticities and the possibility of complementarity in production and consumption. A number of cases arise – such as a 'transfer paradox' in which a transfer immiserizes the recipient country – whose existence is inconsistent with market stability in two-country two-commodity models (on the stability question, see Samuelson 1947, 1971; for the case of more than two countries, see Gale 1974; Bhagwati et al. 1983). Extensive literature reviews can be found in Eaton (1989) and Brakman and van Marrewijk (1998).

Research on the transfer problem since 1980 has moved toward intertemporal issues. The analytical tools used are the representative agent and overlapping generations frameworks in either a two-period or infinite-horizon setting. Two-country analyses (for example, Djajić et al. 1998) have focused on the adjustment of the intertemporal terms of trade (the world interest rate) to a transfer. The two-country, overlapping-generations model has attracted particular interest because the transfer paradox can occur with just two countries, since the competitive equilibrium need not be Pareto efficient (see Galor and Polemarchakis 1987; Haaparanta 1989).

Much recent research concerns the financing of a transfer. As noted by Gavin (1992) and Devereux and Smith (2005), France borrowed an amount equivalent to almost one quarter of its GDP in order to finance its reparations payments to Germany during 1872 and 1873. In contrast with static analyses, intertemporal models allow the initial payment of a transfer to be partially financed by international borrowing (Sachs 1981; Obstfeld and Rogoff 1995). As a simple illustration, consider a small open economy with fixed endowment income $(\bar{y})$ and initial foreign debt (b_{t-1}) that must pay a one-period reparation (τ_t). With a given world interest rate (r), the current account at time t will be $CA_t \equiv \bar{y} - c_t - rb_{t-1} - \tau_t$. Assume perfect foresight and that consumption (c_t) can be approximated by permanent income $(\tilde{y}_t)$:

$$\tilde{y}_t \equiv -rb_{t-1} + \frac{r}{1+r}\left[\overline{y} - \tau_t + \frac{\overline{y} - \tau_{t+1}}{1+r} + \ldots\right]$$
$$\equiv \overline{y} - rb_{t-1} - \frac{r}{1+r}\left[\tau_t + \frac{\tau_{t+1}}{1+r} + \ldots\right],$$

where $\sum_{n=0}^{\infty}\left(\frac{1}{1+r}\right)^n = \frac{1+r}{r}$. Then the current account will be determined by the timing and magnitude of current and future reparations payments:

$$CA_t = -\tau_t + \frac{r}{1+r}\left[\tau_t + \frac{\tau_{t+1}}{1+r} + \ldots\right].$$

Given an initially balanced current account, a reparations payment in period t only will result in a current account deficit of $\left(\frac{1}{1+r}\right)\tau_t$, thereby creating a debt obligation that spreads out the necessary accompanying decline in consumption by $\left(\frac{r}{1+r}\right)\tau_t$ across all periods from time t onward. The imposition of a uniform reparations payment in all time periods $(\tau_t = \tau_{t+1} = \ldots = \overline{\tau})$ will lower consumption by $\overline{\tau}$ so that the improvement in the trade account $(TA_t \equiv \overline{y} - c_t)$ exactly offsets the reparations payment in period t (a condition that is imposed in a static analysis of the transfer problem). In settings involving reproducible capital, reparations payments (or other transfers) may have an additional impact on current account dynamics by setting in motion a gradual adjustment of the capital stock to a new equilibrium (Brock 1996; Chatterjee et al. 2003).

Beginning with Krugman (1999), the term 'transfer problem' has been applied to an abrupt reduction in capital flows during an economic crisis. The formal similarity between adjustment to a capital flow reduction and a reparations payment can be seen from the following trade account identity:

$$TA_t \equiv CA_t + rb_{t-1} + \tau_t$$

where $CA_t \equiv b_{t-1} - b_t$. If we hold the current account constant, an increase in reparations payments $(\tau_t\uparrow)$ requires an accompanying improvement in the trade account. A reduction of capital flows $(b_t\downarrow)$ requires a similar improvement in the current and trade accounts. However, since there is a binding borrowing constraint and no direct wealth effect associated with an abrupt reduction in capital flows, this second transfer problem is conceptually distinct from the classical transfer problem, thus supporting Nurkse's (1961) admonition not to apply indiscriminately the analysis of an unrequited transfer to problems involving international capital movements.

Despite the large literature on the static and intertemporal dimensions of unrequited transfers, there is relatively little empirical research on the transfer problem. Papers by Yano and Nugent (1999), Lane and Milesi-Ferretti (2004), Devereux and Smith (2005), and Rajan and Subramanian (2005) are notable promising exceptions.

See Also

- ▶ Keynes, John Maynard (1883–1946)
- ▶ Ohlin, Bertil Gotthard (1899–1979)
- ▶ Terms of Trade

This is a revision and extension of the article in the first edition of the New Palgrave dictionary by Barry Eichengreen.

Bibliography

Bhagwati, J.M., R.A. Brecher, and T. Hatta. 1983. The generalized theory of transfers and welfare. *American Economic Review* 73: 606–618.

Brakman, S., and C. van Marrewijk. 1998. *The economics of international transfers*. Cambridge, UK: Cambridge University Press.

Brock, P. 1996. International transfers, the relative price of non-traded goods, and the current account. *Canadian Journal of Economics* 29: 163–180.

Chatterjee, S., G. Sakoulis, and S. Turnovsky. 2003. Unilateral capital transfers, public investment, and economic growth. *European Economic Review* 47: 1077–1103.

Devereux, M., and G. Smith. 2005. *Transfer problem dynamics: Macroeconomics of the Franco-Prussian war indemnity*. Working paper no. 1025, Department of Economics Queens University, Kingston.

Djajić, S., S. Lahiri, and P. Raimondos-Moller. 1998. The transfer problem and the intertemporal terms of trade. *Canadian Journal of Economics* 31: 327–336.

Eaton, J. 1989. Foreign public capital flows. In *Handbook of development economics*, ed. H. Chenery and T.N. Srinivasan, vol. 2. Amsterdam: North-Holland.

Eichengreen, B. 1986. Macroeconomics and history. In *The future of economic history*, ed. A. Field. Boston: Martinus Nijhoff.
Gale, D. 1974. Exchange equilibrium and coalitions. *Journal of Mathematical Economics* 1: 63–66.
Galor, O., and H. Polemarchakis. 1987. Intertemporal equilibrium and the transfer paradox. *Review of Economic Studies* 54: 147–156.
Gavin, M. 1992. Intertemporal dimensions of international economic adjustment. *American Economic Review* 82: 174–179.
Haaparanta, P. 1989. The intertemporal effects of international transfers. *Journal of International Economics* 26: 371–382.
Jones, R. 1975. Presumption and the transfer problem. *Journal of International Economics* 5: 263–274.
Keynes, J.M. 1929. The German transfer problem. *Economic Journal* 39: 11–17.
Krugman, P. 1999. Balance sheets, the transfer problem, and financial crises. *International Tax and Public Finance* 6: 459–472.
Lane, P., and G. Milesi-Ferretti. 2004. The transfer problem revisited. *The Review of Economics and Statistics* 86: 841–857.
Mundell, R.A. 2002. Keynes and Ohlin on the transfer problem. In *Bertil Ohlin: A centennial celebration (1899–1999)*, ed. R. Findlay, L. Jonung, and M. Lundahl. Cambridge, MA: MIT Press.
Nurkse, R. 1961. Causes and effects of capital movements (1933). In *Equilibrium and growth in the world economy*, ed. G. Haberler and R. Stern. Cambridge, MA: Harvard University Press.
Obstfeld, M., and K. Rogoff. 1995. The intertemporal approach to the current account. In *Handbook of international economics*, ed. G. Grossman and K. Rogoff, vol. 3. Amsterdam: North-Holland.
Ohlin, B. 1929. The reparation problem: A discussion. *Economic Journal* 39: 172–183.
Pigou, A.C. 1932. The effects of reparations on the ratio of international exchange. *Economic Journal* 42: 532–543.
Rajan, R., and A. Subramanian. 2005. *What undermines aid's impact on economic growth?* Working paper no. 05/126, Washington, DC: International Monetary Fund.
Sachs, J. 1981. The current account and macroeconomic adjustment in the 1970s. *Brookings Papers on Economic Activity* 1981 (1): 201–282.
Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P. 1952. The transfer problem and transport costs. *Economic Journal* 62: 278–304.
Samuelson, P. 1971. On the trail of conventional beliefs about the transfer problem. In *Trade, balance of payments and growth*, ed. J. Bhagwati et al. Amsterdam: North-Holland.
Yano, M., and J. Nugent. 1999. Aid, nontraded goods, and the transfer paradox in small countries. *American Economic Review* 89: 431–449.

Transformation of Statistical Variables

D. R. Cox

Abstract

Transformations of many kinds are used in statistical method and theory including simple changes of unit of measurement to facilitate computation or understanding, and the linear transformations underlying the application and theory of multiple regression and the techniques of classical multivariate analysis. Nevertheless the word transformation in a statistical context normally brings to mind a non-linear transformation (to logs, square roots, etc.) of basic observations done with the objective of simplifying analysis and interpretation. The present entry focuses on that aspect.

Transformations of many kinds are used in statistical method and theory including simple changes of unit of measurement to facilitate computation or understanding, and the linear transformations underlying the application and theory of multiple regression and the techniques of classical multivariate analysis. Nevertheless the word transformation in a statistical context normally brings to mind a non-linear transformation (to logs, square roots, etc.) of basic observations done with the objective of simplifying analysis and interpretation. The present entry focuses on that aspect.

Mostly we discuss problems in which variation in a univariate response variable, y, is to be explained in terms of explanatory variables $x_1,\ldots,x_p$; the terminology here is self-explanatory and avoids overuse of the words dependent and independent! We consider transformations of y and/or some or all of the explanatory variables. Note that where a number of variables are of very similar kinds, it may be sensible to insist on transforming them in the same way.

A brief historical note is desirable. Until the wide availability of computers, the majority of

relatively complicated statistical analyses used the method of least squares or fairly direct elaborations thereof. Particular requirements of these methods are linear representations of the expected response, constancy of variance and normality of distribution of errors. When the data manifestly do not obey one or more of these conditions, transformation of variables provides a flexible and powerful technique for recovering a situation to which well-understood methods of analysis are reasonably applicable and thus greatly extends the range of applicability of those methods. With powerful and sometimes even flexible computing facilities now commonplace, such transformations, while remaining important, are less so than they used to be, because it is now feasible to develop special models for each specific application and to implement an appropriate analysis from first principles.

Purpose of Transformations

The key assumptions of the 'classical' methods mentioned above are (a) simplicity of structure, additivity, linearity, absence of interaction; (b) constancy of variance; (c) normality of error distribution. Independence of errors is another very important assumption, needing especially careful consideration in the case of time series data, but is not particularly relevant in the present discussion.

While the relative importance of (a)–(c) depends on the context, they are listed in broadly decreasing order of importance. Linear relations are easy to specify and understand; absence of interaction, for example that important relations retain their form for different groups of data, is important not only for understanding but also as a basis for extrapolation to new groups of data.

Constancy of variance has a triple role. If the pattern of variance is of intrinsic interest, constancy of variance is a reference level for interpretation. If the effect of explanatory variables on whole distributions is of interest, constancy of variance suggests that only changes in location need be studied. Finally constancy of variance is required for various technical statistical reasons. Appreciable changes in variance vitiate standard errors and tests of significance and will lead to a general loss of efficiency; the method of weighted least squares can be used when the nature of the changes in variance is at least roughly known.

The assumption of normality of error distributions is particularly important if the ultimate objective is prediction in the tails of a distribution. Otherwise appreciable non-normality is sometimes an indication that a quite different distributional formulation is called for, sometimes a warning about the occurrence of aberrant values in the data and more broadly is a sign of potential loss of efficiency and possible failure of tests of significance.

The possibility of approximately satisfying all three requirements simultaneously is often an expression of rational optimism, to be assumed although not taken for granted.

An important aspect of any statistical analysis is the presentation of conclusions in a simple form and this may demand reinterpretation of conclusions of conclusions on to the original scale of measurement.

Construction of Transformations

We now discuss in outline a number of techniques for choosing a suitable transformation.

The two most important techniques are probably previous experience of similar data, and the application of diagnostic checks to the analysis of untransformed data. In the latter case it may be clear that 'pulling in' either of the upper tail or of the lower tail of the data would be helpful.

To stabilize variance, a widely used technique is to establish either empirically or theoretically a relation between variance and mean. If for observations of true mean μ the variance is $v(\mu)$, then it is easy to show by local linearization that the transformation

$$y \to \int_0^y dx/\sqrt{v(x)}$$

will induce observations of approximately unit variance. A common possibility is to find $v(\mu)$

approximately of the form $a\mu^b$, often established by plotting log sample variance against log sample mean, when a line of slope b should result. This leads to a power transformation except for $b = 2$, when a log transformation is indicated. The z transformation of correlation coefficients, r,

$$r \to \frac{1}{2}\log\{(1+r)/(1-r)\}$$

is historically the first example of this argument, the relation between mean and variance being obtained theoretically.

Some simple equations expression non-linear relations have simple linearizing transformations, of which the most common and important is the relation

$$y = \alpha x_1^{\beta 1} x_2^{\beta 2},$$

which is linearized by taking logs of all variables. A more empirical approach, not in fact much used in practice, is to search within some family of possible transformations for one which minimizes a measure of non-linearity or interaction.

A much more formal approach to the choice of a transformation is to start with some parametric family of transformations $y \to y^{(\lambda)}$ of which the most important is normally the family of power transformations, including as a limiting case the log transformation. The unknown parameter λ indexes the transformation that is appropriate. If now it is assumed that for some unknown λ the transformed values satisfy all the standard assumptions of some special convenient model, such as the normal theory general linear model, formal methods of estimation, in particular the method of maximum likelihood, can be applied to estimate λ, to see whether there is evidence that a transformation really does improve fit, to compare the values of λ in several unrelated sets of data, and so on. The calculations are relatively simple and straightforward. The usual procedure is to choose as a scale for analysis that corresponding to a simple value of λ reasonably consistent with the data.

Transformations to normality are always possible for a single continuous distribution, because any continuous distribution can be transformed into any other. Normalizing transformations are quite widely used in theoretical arguments; their direct use in the analysis of empirical data is on the whole rather less common, essentially for the reasons outlined above.

Some Further Developments

The topics outlined above have an extensive literature. Some recent points of discussion and open issues are as follows:

(i) There are no good techniques for the transformation of multivariate distributions other than component by component.

(ii) Transformation selection by methods that are robust to outliers have been discussed, although in many practical situations it is the extreme observations that carry the most information about the appropriateness of transformations and whose accommodation is particularly important.

(iii) Following the choice of a transformation estimation and interpretation of effects is usually carried out on the transformed scale as if this had been given a priori. The appropriateness and justification of this has been the subject of lively discussion.

(iv) It is possible to transform to simple models other than the standard normal ones, for example to the exponential based models so useful in the analysis of duration data.

(v) The main procedures discussed above involve an interpretation essentially in terms of the expected response on the transformed scale. An alternative approach postulates that the expected value of the response on the original scale is a suitable non-linear function of a linear combination of explanatory variables. To distinguish empirically between these formulations is likely to require a large amount of high quality data.

(vi) Methods can be developed for estimating transforming functions totally non-parametrically. Such an approach uses a great deal of computer time.

See Also

- ▶ Non-linear Methods in Econometrics
- ▶ Regression and Correlation Analysis

Bibliography

Bartlett (1943) gives an excellent account of the early work; Box and Cox (1964) discuss the estimation of transformations via the likelihood and Bayesian methods. Butter and Verbon (1982) describe economic applications in some depth. Bickel and Doksum (1981) and Box and Cox (1982) give opposing views of estimation following a transformation.

Bartlett, M.S. 1947. The use of transformations. *Biometrics* 3: 39–52.

Bickel, P.J., and K.A. Doksum. 1981. An analysis of transformations revisited. *Journal of the American Statistical Association* 76: 296–311.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26: 211–243.

Box, G.E.P., and D.R. Cox. 1982. An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* 77: 209–210.

den Butter, F.A.G., and H.A.A. Verbon. 1982. The specification problem in regression analysis. *International Statistical Review* 50: 267–283.

Transformation of Variables in Econometrics

Paul Zarembka

Economic theory usually fails to describe the functional relationship between variables (the CES production function being an exception). In econometrics, implications of simplistic choice of functional form include the danger of misspecification and its attendant biases in assessing magnitudes of effects and statistical significance of results. It is safe to say that when functional form is specified in a restrictive manner a priori before estimation, most empirical results that have been debated in the professional literature would have had a modified, even opposite, conclusion if the functional relationship had not been restrictive (see Zarembka 1968, p. 509, for an illustration; also, Spitzer 1976).

Most econometric research is not based on a large enough sample size for elaborate functional relationships to be meaningful. Therefore, a functional relationship which preserves additivity of effect (as in the linear model), but is more general than the usual choice between linear and linear-in-logarithmic models, ought to be sufficiently general. A transformation of variables in the form

$$y^{(\lambda)} = \begin{cases} \left(y^{\lambda} - 1\right)/\lambda, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

offers a solution, where we note that $\lim \lambda \to 0$ in the upper expression is in fact $\ln y$, the lower expression.

Such transformations can be applied both to the dependent and independent variables with additivity of effect preserved on the transformed variable(s). The linear and linear-in-logarithmic models are both special cases ($\lambda = 1$ and $\lambda = 0$ for all variables, respectively). The transformation on the dependent variable may be different from those for the independent variables and different transformations may be applied to different independent variables, with corresponding increases in the parameter space (an obvious extension not elaborated upon here). It is important to note that a constant term must be included as an independent variable in order to preserve invariance of estimates of a transformation on a variable to changes in units of measurement for that variable (Schlesselman 1971); otherwise the form y^{λ} or y^{λ}/λ would have to be used.

Usual econometric practice with a linear model, in this case in transformed variables, is to add an error term with a normal distribution of zero mean and constant variance (actually, here, only approximately normal since negative values on a transformed dependent variable would generally be truncated). Using maximum likelihood estimation, Box and Cox (1964) follow this approach. However, Amemiya and Powell (1981) have questioned such a procedure on grounds that the error distribution must be truncated and they show with the example of a gamma

distribution on the dependent variable before transformation that the Box and Cox procedure is inconsistent and typically leads to quite different results than their statistically proper procedure. Nevertheless, Draper and Cox (1969) have shown that, as long as the error term is reasonably symmetric, the Box and Cox procedure is robust. (The apparent discrepancy between Draper and Cox and Amemiya and Powell is presumably that the latters' assumed gamma distribution on the untransformed dependent variable need not imply a 'reasonably' symmetric distribution of the transformed variable. The issue deserves further research.)

Following usual practice in its assumed normality (albeit here truncated) of the error term, along with independent and identically distributed terms, the iterated ordinary least squares is conceptually the simplest procedure for estimation. If σ^2 is the constant variance of the error distribution and N is the number of observations, then the maximized log-likelihood L for given λ is, except for a constant,

$$-\frac{1}{2}N\ln\hat{\sigma}^2(\lambda)+(\lambda-1)\sum_{i-1}^{N}\ln y_i,$$

or, with the specific sample of y scaled by its geometric mean so that the latter term is zero,

$$L_{\max}(\lambda)=-\frac{1}{2}N\ln\hat{\sigma}^2(\lambda).$$

To maximize over the whole parameter space, simply take alternative values of λ; the one that minimizes $\hat{\sigma}^2(\lambda)$ maximizes the log likelihood (a procedure almost any simple least squares programme can handle). Estimates of parameters for independent variables are also thus provided, while (as emphasized by Spitzer 1982a, p. 311) their standard errors should be obtained from the information matrix. An approximate $100(1-\alpha)$ per cent confidence region for λ can be obtained from

$$L_{\max}\left(\hat{\lambda}\right)-L_{\max}(\lambda)<\frac{1}{2}\chi_1^2(\alpha).$$

For example, for a 95 per cent confidence interval ($\alpha=0.05$), the region can be obtained from $L_{\max}\left(\hat{\lambda}\right)-L_{\max}(\lambda)<1.92$. (Beauchamp and Kane (1984), give a survey of other estimation procedures, including Spitzer's (1982b) modified Newton algorithm preferred by him for its greater computational efficiency.)

The above has assumed that the error distribution is homoskedastic across observations on the dependent variable as transformed by the true λ. While such an assumption is relaxingly convenient, there is no obvious justification for it. Zarembka (1974, pp. 87–95) has analysed the circumstance in which heteroskedasticity of the error term obtains and shows that an incorrect assumption of homoskedasticity implies that the resulting maximum-likelihood procedure is biased asymptotically away from the true λ toward that transformation which more nearly stabilizes the error variance. Other parameters will also fail to be consistently estimated. For example, if the variance of the untransformed dependent variable y_i is constant, the bias is toward $\lambda=1$; if the coefficient of variation of y_i is constant, the bias is toward $\lambda=0$.

To estimate consistently the model under heteroskedasticity, some assumption concerning its pattern seems required. Zarembka (1974, pp. 93–5) considers the case where the variance of y_i is related to the power of the expectation of the transformed y_i Lahiri and Egy (1981) consider the case where the variance of the transformed y_i is related to a power of any exogenous variable while Gaudry and Dagenais (1979) relate that variance to a function of several exogenous variables. Seaks and Layson (1983) consider the case where the variance of the transformed dependent variable is related to the square of one of the independent variables, while they also include autocorrelation in the error terms as an additional possibility (a problem first tackled for such transformation-of-variables models by Savin and White 1978; see also Gaudry and Dagenais 1979). Most show through examples the actual importance of confronting the possibility of heteroskedasticity. In the absence of specifying a structure for heteroskedasticity, Tse (1984) suggests a Lagrange multiplier test for its possible presence.

Almost all empirical uses of transformation of variables still use the conventional assumptions of normally and independently distributed error terms of constant variance (with exceptions such as Blaylock and Smallwood 1982). In econometrics, examples of the wide range of applications of transformations of variables have included the elasticity of factor substitution in neoclassical production economics (and its possible variability), economies of scale in banking, the rate of technical progress represented in the 'Indianapolis 500', willingness to pay for automobile efficiency, economic depreciation of used buildings, capital asset pricing models, elasticities of demand for consumption (and specifically, meat, food, radios, and air quality), elasticities of the demand for money (and a possible 'liquidity trap') and for imports, demand for leisure and non-pecuniary job characteristics, relation of earnings to schooling and cognitive abilities, wage and rent gradients and the price elasticity of demand for urban housing, and elasticities of interstate migration.

See Also

- ▶ Non-linear Methods in Econometrics
- ▶ Regression and Correlation Analysis

Bibliography

Amemiya, T., and J.L. Powell. 1981. A comparison of the Box–Cox maximum likelihood estimator and the non-linear two-stage least squares estimator. *Journal of Econometrics* 17: 351–381.

Beauchamp, J.J., and V.E. Kane. 1984. Application of the power-shift transformation. *Journal of Statistical Computation and Simulation* 19: 35–58.

Blaylock, J.R., and D.M. Smallwood. 1982. Analysis of income and food expenditure distributions: A flexible approach. *Review of Economics and Statistics* 64: 104–109.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26: 211–243.

Draper, N.R., and D.R. Cox. 1969. On distributions and their transformation to normality. *Journal of the Royal Statistical Society, Series B* 31: 472–476.

Gaudry, M.J.I., and M.G. Dagenais. 1979. Heteroskedasticity and the use of Box-Cox transformations. *Economics Letters* 2: 225–229.

Lahiri, K., and D. Egy. 1981. Joint estimation and testing for functional form and heteroskedasticity. *Journal of Econometrics* 15: 299–307.

Savin, N.E., and K.J. White. 1978. Estimation and testing for functional form and autocorrelation. *Journal of Econometrics* 8: 1–12.

Schlesselman, J. 1971. Power families: a note on the Box and Cox transformation. *Journal of the Royal Statistical Society, Series B* 33: 307–311.

Seaks, T.G., and S.K. Layson. 1983. Box–Cox estimation with standard econometric problems. *Review of Economics and Statistics* 65: 160–164.

Spitzer, J.J. 1976. The demand for money, the liquidity trap, and functional forms. *International Economic Review* 17: 220–227.

Spitzer, J.J. 1982a. A primer on Box–Cox estimation. *Review of Economics and Statistics* 64: 307–313.

Spitzer, J.J. 1982b. A fast and efficient algorithm for the estimation of parameters in models with the Box–Cox transformation. *Journal of the American Statistical Association* 77: 760–766.

Tse, Y.K. 1984. Testing for linear and log-linear regressions with heteroskedasticity. *Economics Letters* 16: 63–69.

Zarembka, P. 1968. Functional form in the demand for money. *Journal of the American Statistical Association* 63: 502–511.

Zarembka, P. 1974. Transformation of variables in econometrics. In *Frontiers in econometrics*, ed. P. Zarembka. New York: Academic Press.

Transformation Problem

E. K. Hunt and Mark Glick

The 'transformation problem' is at the heart of the Marxian labour theory of value. The topic has always been the subject of sharp controversy. The controversy reflects not only the general ideological conflicts that surround all Marxist ideas, but also the disagreement among Marxists themselves about the nature of the labour theory of value. After defining the problem in Marx's terms, we first present Marx's solution and the claims which he makes regarding its properties. This discussion is then followed by a brief critical review of the various solutions which have been proposed since Marx.

The 'Transformation Problem'

To Marx, the value of a commodity consisted of the labour embodied in the means of production that were used up in the production of the commodity (dead labour) and the labour expended in the current production period (living labour).

$$W = L_{\mathrm{d}} + L_{\mathrm{l}} \tag{1}$$

where W is value, L_{d} is dead labour, and L_{l} is living labour. Living labour can be separated into necessary labour L_{n} and surplus labour, L_{s}. Necessary labour is that proportion of living labour that creates the value equivalent of the worker's wages and surplus labour is the remaining living labour time during which the value equivalent of surplus value is created. Thus, the following equation holds:

$$W = L_{\mathrm{d}} + L_{\mathrm{n}} + L_{\mathrm{s}} \tag{2}$$

In actual pricing processes, Marx believed that capitalists summed up the costs of production and then added a percentage markup, which was determined by the average rate of profit. Thus the formula for equilibrium prices is:

Price of Prod = cost of commodities + cost of labour + Profit markup or, using p for the prices of production, c for constant capital, v for variable capital, and r for the rate of profit, we have:

$$P = c + \upsilon + r(c + \upsilon)$$

where $r + s/c + v$ and $r(c + v) = s/c + v$ $(c + v) = s$.

The general correspondence between the various types of labour and the cost-components of price is obvious:

$$\begin{array}{ccccccc} W & = & L_d & + & L_n & + & L_s \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ P & = & c & + & \upsilon & + & r(c+\upsilon). \end{array}$$

Price corresponds to value, constant capital corresponds to dead labour; variable capital corresponds to necessary labour; and profit corresponds to surplus value.

The most important reason why this correspondence is not proportional or one-to-one, however, is that the production of different commodities involves unequal organic compositions of capital (defined as either c/v or $L_{\mathrm{d}}/L_{\mathrm{l}}$). The exchange of commodities at values is thus incompatible with equal rates of profit. Given two industries which exchange at values, their rates of profit can only be equal if their organic compositions are equal:

$$r_1 = \frac{s_1\upsilon_1}{c_1/\upsilon_1 + 1} = \frac{s_2/\upsilon_2}{c_2/\upsilon_2 + 1}.$$

Since the rates of surplus will be equalized through competition between workers, it follows that equal rates of profit imply equal organic compositions $(c_1/v_1) = (c_2/v_2)$. Marx argued that this would not be the case in general and that (c/v) varied significantly from sector to sector.

Marx's Solution

Marx's solution to the problem was to transform values into prices of production which correspond to equalized rates of profit. In chapter 9 of Volume III of *Capital* he presents a table with five sectors and transforms values to prices of production by the following procedure. First he calculates the average rate of profit as:

$$\frac{\sum S_i}{\sum (c_i + \upsilon_i)}$$

Once given the average rate of profit, r, he recalculates all of the prices according to the formula:

$$(1 + r)(c_i + \upsilon_i)$$

Marx was anxious to show that the essence of the labour theory of value and the theory of surplus value can be preserved when the transition is made from values to prices. Prices, he argued, are merely transformed values and profit is redistributed surplus value. In order to show the consistency of this view he made the following two claims concerning the aggregates in his

transformation solution: 1. The sum of values = the sum of prices; 2. The sum of surplus value = the sum of profit.

> The sum of the profits for all the different spheres of production must accordingly be equal to the sum of surplus values, and the sum of prices of production for the total social product must be equal to the sum of its values (Marx [1867], 1981, p. 273).

The equality of these aggregates was used by Marx to argue that only a redistribution has occurred and nothing has actually been created or destroyed in the transformation from values to prices.

The Reproduction Scheme Argument: Bortkiewicz, Sweezy, Seton

Following the publication of Marx's solution to the transformation problem, a number of critics pointed out that Marx had not completely solved the transformation problem. In his solution, Marx had transformed the output prices while the input prices remained in values. This was an inadequate solution, it was argued, since capitalists buy inputs at prices and not values. In addition, the output price of one commodity is the input price of another. In his famous 1907 article, Bortkiewicz attempted to solve the transformation problem by simultaneously transforming both inputs and outputs (Bortkiewicz 1907). But in his result he found that he could obtain only one of the two claims made by Marx. Either total prices were equal to total values or total surplus value was equivalent to total profit, but not both. He considered this as an important criticism of the labour theory of value.

In 1942, Sweezy built on the Bortkiewicz result using a three sector reproduction scheme. Although Bortkiewicz used this apparatus as a matter of convenience, Sweezy argued that the transformation procedure should 'not result in a disruption of the conditions of simple reproduction' (Sweezy 1942, p. 114). Sweezy went beyond Bortkiewicz, and claimed that his solution would satisfy both of Marx's claims. He obtained such a result by assuming that the output of the luxury sector is equal to unity, and assuming that this sector also has the average organic composition. He argued that these two assumptions are reasonable since the output of the luxury sector can be considered the money commodity, and to avoid price/value deviations in the money commodity, its organic composition must be set equal to the average of the first two sectors. Seton later provided a proof of Sweezy's example.

Unfortunately, Sweezy's success is a result of his assumptions. First, since surplus value is equal to the output of the luxury sector, setting this output equal to one in both prices and values ensures that total surplus value will equal total profit. The assumption of a socially average organic composition in the third sector obtains the second condition. If the sum of the organic compositions of department I and department II is equal to that of department III, and department III's output is set equal in prices and values, then the sum of prices and values in departments I and II must also be equal. Not only are Sweezy's results true by definition, but these two assumptions are unnecessarily restrictive for a convincing solution to the transformation problem.

Normalization by Sraffa's Standard Commodity: Medio

In general, it can be said that, lacking an invariant measure of value, it has proven impossible to obtain a transformation solution in which the equalities between values and prices as well as profit and surplus value can be simultaneously maintained without the aid of extremely restrictive assumptions. When Sraffa's standard commodity became widely known there was initially some hope that it might provide such an invariant measure. This hope was quickly abandoned, however, when it was realized how restrictive the nature of the invariance of Sraffa's standard commodity is.

Marx, however, suggested a third method for linking prices to labour values. It is within the context of this third method that Alfredo Medio (1972) demonstrated that Sraffa's standard commodity could provide an important analytical tool for the Marxist labour theory of value. Marx realized that if a commodity could be found that was produced with the socially average organic

composition of capital, then the rate of profit which could be obtained in the production and sale of that commodity would be identical whether all commodities were sold at their labour values or at their transformed money prices. Therefore, the rate of profit on that commodity would be determined entirely by labour values. Moreover, since competition tended to equalize all profit rates, it could be shown that the socially average rate of profit (by virtue of which all price calculations could be made with a cost-of-production theory of prices) would correspond to the rate of profit on the average commodity – a rate determined entirely by labour value calculations. If a numeraire that equates aggregate profit and aggregate surplus value (or equates the aggregate of values and prices) cannot be found, then an average industry whose rate of profit is determined by labour values suffices to connect the labour value analysis and the price analysis.

Medio demonstrated that in the industry producing Sraffa's standard commodity, the Marxian formula for the rate of profit, $r = (s/v)/(c/v + 1)$, always holds true. In Medio's demonstration the profit rate (r) is the money rate of profit by which capitalists mark up their money costs to arrive at prices. The rate of exploitation, or rate of surplus value (s/v), is defined in labour value terms. It is the rate at which surplus value is created in the sphere of production, and hence it is equal in all industries. The organic composition of capital (c/v), however, has a special meaning in Medio's formulation. It is determined by labour values alone, and is a weighted average of all of the production processes that make up the industry that produces the standard commodity.

Medio's solution has been criticized by the observation that a standard commodity does not actually exist, and that a hypothetical form of measurement is a weaker claim than that sought by Marx.

The Iterative Method and Balanced Growth: Shaikh

Anwar Shaikh's popular solution to the transformation problem has been published in two important papers with a seven year gap (Shaikh 1977, 1984). In his 1977 paper on the transformation problem, Shaikh is concerned with establishing a link between Marx's method and what he considers the 'correct' prices obtained by Bortkiewicz. Instead of developing a new mathematical apparatus, all one had to do, according to Shaikh, is to iterate Marx's procedure. If one takes Marx's prices of production and uses them as inputs, and then uses Marx's procedure again to obtain new prices of production, and so on, one converges on the set of Bortkiewicz prices. Shaikh's actual procedure, however, makes a number of assumptions which are found in Bortkiewicz but may not be in Marx. He sets the sum of prices equal to the sum of values in each step, and adjusts the money wage at every step so that the workers consume a certain bundle of commodities at the previous period's prices. Shaikh's procedure does obtain the set of prices consistent with the Bortkiewicz method, but also like Bortkiewicz, he obtains only one of Marx's aggregates. In Shaikh's solution total surplus value is not equal to total profit. Why not? This is the issue discussed in his 1984 paper.

In his 1984 paper, Shaikh argues that the transformation solution should not adopt ad hoc assumptions to obtain both of Marx's aggregates. Instead, he reasons, we should actually expect total surplus value and total profit to differ. This difference is due to the price-value deviations and the size of the luxury sector. When price-value deviations exist in the luxury sector, surplus value can be gained or lost through the circuits of revenue. His proof of this argument utilizes the assumption of balanced growth. In a situation of balanced growth he shows that the difference between surplus value and profit can be shown to be proportional to the price-value deviation in the sector producing luxury products. Such a result is very close to the well known property of von Neumann systems that when an economy is at maximum balanced growth and one of Marx's claims is assumed, then the other will automatically follow. Unfortunately, Shaikh's result cannot hold in a real economy where balanced growth is not satisfied.

The 'New Solution': Duménil, Lipietz and Foley

What is being called the 'new solution' to the transformation problem by a small but growing group of Marxist economists was first introduced to English-speaking readers by Lipietz (1982). but the original solution was formulated by Duménil (1980) and later 'discovered' independently by Duncan Foley (1982). The new solution entails two important assumptions which are traced back to Marx. The first is that (the sum of prices equals the sum of values) should be modified to read: the sum of the prices of the net product (defined as the value added) should be the sum of the values of the net product. The second assumption is that distribution must be defined ex post, as either the value of the money wage which workers receive (Foley 1982), or the bundle of consumption goods which the workers buy valued at prices (Duménil 1980). Once these two assumptions are made any set of values can be transformed into any set of prices with the property that both of Marx's aggregates hold.

Duménil and Foley make two arguments for the adoption of their unique normalization procedure on the net product. First, they claim that such a normalization avoids double counting (Duménil 1983–4, p. 442). In addition, they both argue that such a normalization conforms to Marx's view of what value is. Value 'is the linking of the total labour expended in a given period with the production associated with it, that is, the net product' (Duménil 1983–4, p. 442). In addition, they argue that wages must be evaluated on the basis of prices and not as the value of a wage bundle. This view of distribution avoids the problem that when prices deviate from values, the rate of exploitation in price terms depends on the particular set of goods which workers buy and is not settled in the production process. They further argue that, in the previous formulations, if any part of the wage is saved the rate of surplus value becomes incalculable. Foley goes further than Duménil and argues that the wage should not be considered as a bundle at all. Wages are a sum of money, he claims, which can be used to buy any goods at the existing set of prices. In addition, unlike a wage bundle, the money wage conceals the exploitative nature of capitalist relations (Foley 1982, p. 43).

One argument which has been posed against this view is that in the set of 'new solution' prices of production the sum of the values of constant capital does not equal the total sum of its prices. A convincing argument justifying this result must be established. In addition, the distribution assumption requires ex post knowledge. The actual set of prices must be known before the rate of wages can be established. One cannot move step by step from values into prices. The two realms must be considered separately while the new solution only provides a mapping procedure from one to the other.

Summary and Implications

The transformation problem arose from the attempt to show that the labour theory of value is consistent with the money prices of exchange. Marx's two claims that total prices should be equal to total values and total surplus value should be equal to total profit have traditionally been considered a prerequisite to the argument that prices are merely transformed values and profit is redistributed surplus value. We have shown that this result can be obtained by using numerous different assumptions. Some of these procedures hold total prices and values constant but require special assumptions to obtain an equality between surplus value and profit, others do the reverse. Many of these assumptions are clearly unjustifiable while others are rather more realistic. The 'new solution' of Duménil and Foley obtains both aggregates but finds a discrepancy between constant capital in price and value terms, while Medio's solution holds that equality of the rate of profit in value and money terms is more important than either of the two more traditional equalities.

It is clear from the literature on the 'transformation problem' that its resolution will not be merely a mathematical exercise. The ground of this continuing debate in the future will, instead, concern the social and economic implications of the competing assumptions which are adopted and their compatibility with the tenets of the labour theory of value. This, however, will be a complex debate since Marxists themselves have strong disagreements about the specific nature of the labour theory of value as well as its role or function within the Marxist theoretical system.

See Also

- ▶ Bortkiewicz, Ladislaus von (1868–1931)
- ▶ Sweezy, Paul Marlor (1910–2004)
- ▶ Value and Price

Bibliography

Bortkiewicz, L. 1907. Value and price in the Marxian system. Trans. in *International Economic Papers* No. 2, 1952, 5–61.

Duménil, G. 1980. *De la valeur aux prix de production*. Paris: Economica.

Duménil, G. 1983–4. Beyond the transformation riddle: A labor theory of value. *Science and Society* 47(4): 427–450.

Foley, D. 1982. The value of money, the value of labor power and the Marxian transformation problem. *Review of Radical Political Economics* 14(2): 37–47.

Lipietz, A. 1982. The so-called 'transformation problem' revisited. *Journal of Economic Theory* 26(1): 59–88.

Marx, K. 1867. *Capital*, vol. I. Harmondsworth: Penguin, 1981.

Medio, A. 1972. Profits and surplus value: Appeareance and reality in capitalist production. In *A critique of economic theory*, ed. E.K. Hunt and J. Schwartz. New York: Penguin.

Seton, F. 1957. The 'transformation problem'. *Review of Economic Studies* 24: 149–160.

Shaikh, A. 1977. Marx's theory of value and the 'transformation problem'. In *The Subtle anatomy of capitalism*, ed. J. Schwartz. Santa Monica: Goodyear.

Shaikh, A. 1984. The transformation from Marx to Sraffa. In *Ricardo, Marx, Sraffa*, ed. E. Mandel. London: Verso.

Sweezy, P. 1942. *The theory of capitalist development*. New York: Monthly Review Press, ch. 7.

Transformations and Invariance

Dale K. Osborne

Theories of measurement have applications throughout economics. Some applications are familiar because they are firmly established in the literature (think of utility theory and price indexes). But some are yet to be incorporated into the wider literature and many potential applications remain to be made. This entry does not survey the applications or the theories (see Pfanzagl 1968; Krantz et al. 1971 on theories of measurement), but (1) attempts to explain a certain kind of invariance principle and (2) shows how the principle can be applied to economic analysis.

The invariance principle can be expressed somewhat loosely like this: a relation S between the numbers that represent measures of things is interesting (in the sense that it possibly represents a relation between the things themselves) only so far as it is invariant under all permissible changes in the scales of measurement of the things. If a relation S meets this condition, we may reasonably hope that it expresses in the language of numbers a statement that is true of the world; if not, we have no such hope, for the failure of S to hold under all permissible selections of scales shows that it is a property of our analysis of the world, not of the world itself, because the selection of measurement scales is intrinsically arbitrary.

This principle is rarely discussed in economics but it lies behind a number of familiar propositions. Modern economists know that diminishing marginal utility is not an interesting property of an ordinal utility function because it is not invariant under increasing transformations of the utility function. And they compare elasticities, not slopes, of demand functions because order relations among the elasticities, but not among the slopes, are invariant under changes in the

measurement scales for prices and commodities. This invariance does not prove that corresponding order relations exist among the demand functions of the real world, but it allows us to hope so.

By demand functions of the real world, I mean functional relations between commodities and prices, not between their measurements. I do not take the extreme operationalist view that all empirical laws are simple relations between measurements. This view seems untenable if for no other reason than the many relations that surely predated the development of the number systems: preference relations, production functions, hierarchical relations, and so on through a long list. Popper (1963) gives additional objections to operationalism.

Extreme operationalism is one of several conceptual schemes, or schema, for thinking about the relations among measurement, empirical laws, and numerical laws. It is the schema that recognizes no distinction between empirical laws (true statements about the world) and numerical laws (true statements about the numbers that measure the world's things). Three alternative schema are explained in the next section.

Usage Throughout the entry, R denotes the set of real numbers (reals), R_+ the non-negative reals, and R_{++} the positive reals. A function f: $A \rightarrow B$ has a domain in A and a range in B. If the domain is A, the function is *on A* (otherwise *from A*). If the range is B, the function is *onto B* (otherwise *into B*). 'From' and 'into' are understood; 'on' and 'onto' must be stated.

Things and Their Measures

The basic objects of economic behaviour belong to a different category of thought than the numbers that represent them, and relations among these objects belong to a different category than relations among numbers. The goal of economic analysis is to understand the basic objects and the economically interesting relations among them, but this goal is pursued by analysing relations among numbers. Economics is *about* two categories – basic objects and their relations – but it is *performed on* two other categories – numbers and their relations. These two pairs of distinct categories are linked by a fifth category, consisting of the scales of measurement of the basic objects. The linkage is not one-to-one, however, for the scales are not unique.

The distinctions among these categories are central to the part of measurement theory that deals with meaningfulness (a subject discussed by the authors cited above and by many others including Ramsay (1976) and Falmagne and Narens (1983)). The distinctions may be illustrated by an example in which the basic objects are three types of commodities and the relation of interest among them is production. The classic works in production theory verbally describe a two-input–one-output production function as a specified kind of relation between two input commodities and one output commodity, but they formally define it as a real-valued function of two real variables that obeys certain axioms (see Shephard 1970, for example). This equivocation may be explained by a schema. All three of the schema to be mentioned have two elements in common: first, the sets C_k consisting of all commodities of type k ($k = 1, 2, 3$), and second, a function π: $C_1 \times C_2 \rightarrow C_3$ showing which commodity of type 3 is producible by a bundle of the other two.

The most completely specified (that is, least general) schema is that of Dimensional Analysis. In this schema, commodities are regarded as physical quantities in the sense of Whitney (1968) or the more general sense of Krantz et al. (1971), thus paving the way for an application of Luce's (1978) theorem on meaningfulness to π. Then π is a dimensionally invariant relation and may be analysed with the tools of Dimensional Analysis (Bridgman 1922; Kurth 1965; de Jong 1967; Krantz et al. 1971). These tools, familiar to all physicists and engineers, are virtually unknown among modern economists. Jevons's prelude to a rigorous theory of economic dimensions in the second edition of *The Theory of Political Economy* (1879, 1965) was advanced a little by Wicksteed in Palgrave's *Dictionary of Political Economy* and then allowed to die. De Jong's attempted revival (1967) was ignored. Economists just lost interests. Whether this says

something about economists or about economics is far from clear. What is clear is that until commodities are rigorously examined in terms of the axioms of physical quantities, the Dimensional Analytic schema will remain problematic for economics.

A more general schema, which subsumes Dimensional analysis, is that of Conjoint Measurement (Luce and Tukey 1964; Krantz et al. 1971; Luce and Cohen 1983), which characterizes π and the C_k as a unified object of thought (C_1, C_2, C_3, π) and obtains numerical representations of the C_k and π simultaneously. Although this powerful and rapidly developing theory seems naturally suited to economics, its potential applications remain largely unexplored outside the area of decisions under risk.

Although we do not employ an explicit measurement schema in formal economic analysis, our customary procedures embody one implicitly. When we speak of a (two-input, one-output) production function as a real-valued function of two real variables, we have implicitly introduced scales of measurement for the commodities and represented the production function π by what Falmagne and Narens (1983) call a 'numerical code'. Let us make this schema more explicit under the strong assumption that all the measurement scales are ratio scales (unique up to a positive multiplier).

A measurement scale for the type-k commodity is a homomorphism between C_k and R_+, that is, it preserves the algebraic structure of C_k. In assuming ratio scales, we are attributing so much structure to C_k that its homomorphisms are severely limited. The validity of such an attribution is a promising subject for research; here we take it for granted.

When spelled out in sufficient detail to yield an application to production theory, our schema consists of six elements. (1) the sets C_k of commodities of type k. (2) The assumed production function π: $C_1 \times C_2 \rightarrow C_3$. (3) A clear distinction between commodities and numbers. (4) The assumption that, nevertheless, commodities share sufficiently many properties of the real numbers that their homomorphisms, or scales, are determined up to a positive multiplication. The set of scales for type-k commodities is H_k, with members h_k, h'_k, ...; the cartesian product $H_1 \times H_2 \times H_3$ is written as H. We need not decide at this point whether all members of H are available in a system of coherent units. (5) The representation of π by a *general numerical code* F: $R_+ \times R_+ \times F \rightarrow R_+$ defined under the conditions

$$x_k = h_k(c_k), \quad k = 1, 2, 3 \tag{1}$$

$$c_3 = \pi(c_1, c_2) \tag{2}$$

by

$$x_3 = F(x_1, x_2; h_1, h_2, h_3). \tag{3}$$

This representation is general because it holds for all h_k that yield coherent units (i.e. all 'admissible' h_k). It does not appear in practice, where we employ the sixth element. (6) The selection of an arbitrary admissible member h from H and the representation of F (and hence π) by a *special numerical code* f_h: $R_+ \times R_+ \rightarrow R_+$ defined by

$$f_h(x_1, x_2) = F(x_1, x_2; h_1, h_2, h_3), \tag{4}$$

which holds only for the scales h_1, h_2, h_3.

The sixth element of our schema actually consists of a set of special numerical codes like (4), one for each admissible triple of scales. Since every such triple can be expressed in terms of the fixed triple (h_1, h_2, h_3) by (r_1h_1, r_2h_2, r_3h_3) for some positive numbers r_1, r_2, r_3, and every measure x_k, in the fixed scale h_k becomes r_kx_k in the scale r_kh_k, we can write any special numerical code in the generic form

$$f_{rh}(r_1x_1, r_2x_2) = F(r_1x_1, r_2x_2; r_1h_1, r_2h_2, r_3h_3). \tag{5}$$

In practice, of course, the subscripts h and rh are omitted from the functional symbol f.

This schema is one way of making explicit the assumptions and conventions that precede – and link to the world – a typical economic analysis of production, which focuses on a special numerical expression of π.

The preceding references to commodities, ratio scales, and production are only intended to fix ideas. When generalized in the obvious manner, the schema covers all the relations between basic objects that can be represented by functions whose domains and ranges are vectors of real numbers. It leads directly to the following question and sometimes to its answer: What can we learn about f_h and f_{rh}, and thus indirectly about π, from the fact that both f_h and f_{rh} represent π? To answer this question we must know the type of scale in each H_k and something about the relations (if any) among the H_k. In many cases, such knowledge is sufficient for the discovery of interesting properties of π or of unexpected implications of our formal theory of π. In other words, we can learn something about the relations among things – or at least about our theories of such relations – by considering how the things are measured. This remarkable fact ought to be known more widely. We demonstrate its utility by means of the special case of our schema as outlined above.

An Invariance Condition

For national simplicity we may omit the subscript h from the functions f_h and f_{rh} defined in Eqs. (4) and (5) (we are keeping h fixed), denoting them by f and f_r. Since these functions represent the same production function π, we have

$$f(x_1, x_2) = h_3(c_3) = h_3[\pi(c_1, c_2)] \tag{6}$$

$$f_r(r_1x_1, r_2x_2) = r_3h_3(c_3) = r_3h_3[\pi(c_1, c_2)]. \tag{7}$$

If $f(x_1, x_2) > 0$, Eqs. (6) and (7) imply

$$\frac{f_r(r_1x_1, r_2x_2)}{f(x_1, x_2)} = r_3. \tag{8}$$

Clearly, for all (x_1, x_2), (y_1, y_2), …, (z_1, z_2) representing input bundles associated with positive outputs, we have

$$\frac{f_r(r_1x_1, r_2x_2)}{f(x_1, x_2)} = \frac{f_r(r_1y_1, r_2y_2)}{f(y_1, y_2)} = \cdots = \frac{f_r(r_1z_1, r_2z_2)}{f(z_1, z_2)}, \tag{9}$$

showing that the ratio r_3 depends only on r_1 and r_2. Denoting this dependence by ϕ (r_1, r_2), we obtain the equation

$$f_r(r_1x_1, r_2x_2) = \phi(r_1, r_2)f(x_1, x_2). \tag{10}$$

Equation (10) is an invariance condition. It states that the special numerical codes f and f_r of π are related by a positive multiplier that depends only on the multipilers r_1 and r_2. Therefore, properties expressed in terms of the 'internal structure' of f are invariant under changes of the scales for commodities 1 and 2. (See Leontief 1947, for a discussion of 'internal structure'; it is a part of the theory of functional equations, for which Aczel 1966, is a valuable reference.) Equation (10) embodies the invariance principle stated at the beginning of this entry.

Equation (10) is a functional equation in the three functions f_r, φ and f. Given that any two of these functions have positive values at (1, 1), Eq. (10) implies three additional functional equations, one for each function, and all three equations have the same form. The key steps in obtaining these equations use the relations (similar to relation (9)),

$$\frac{f_{sr}(s_1r_1x_1, s_2r_2x_2)}{f(x_1, x_2)} = \cdots = \frac{f_{sr}(s_1r_1z_1, s_2r_2z_2)}{f(z_1, z_2)} \equiv \psi(s_1r_1, s_2r_2)$$

and

$$\frac{f_{sr}(s_1r_1x_1, s_2r_2x_2)}{f_r(r_1x_1, r_2x_2)} = \cdots = \frac{f_{sr}(s_1r_1z_1, s_2r_2z_2)}{f_r(r_1z_1, r_2z_2)} \equiv \epsilon(s_1, s_2).$$

These relations follow because the transformations $x_k \to s_kr_kx_k$ can be decomposed into $x_k \to r_kx_k \to s_kr_kx_k$. Using the definitions of φ, ψ, and ϵ, it can be shown under relatively mild conditions that f satisfies the functional equation

$$f(t_1x_1, t_2x_2) = \frac{f(t_1, t_2)f(x_1, x_2)}{f(1, 1)}. \tag{11}$$

Equation (11) is a stronger invariance condition. It may be interpreted two ways. On the first, t_1 and t_2 represent scale transformations for a fixed input bundle

$$(c_1, c_2) = \left[h_1^{-1}(x_1), h_2^{-1}(x_2)\right],$$

and we see that these transformations affect f only by multiplying all its values by a positive constant $f(t_1, t_2)/f(1, 1)$. On this interpretation, (11) holds for all $(x_1, x_2) \in h_1(C_1) \times h_2(C_2)$ such that $f(x_1, x_2) > 0$ but only for all *admissible* t_1, t_2. On the second interpretation, t_1 and t_2 represent variations in input bundles as measured on a fixed scale, so that $(t_1x_1, t_2x_2) = [h_1(t_1c_1, h_2(t_2c_2)]$. On this interpretation, (11) holds for the same bundles (x_1, x_2) as before and for all (t_1x_1, t_2x_2) such that (t_1c_1, t_2c_2) is defined and productive.

We may thus draw a general conclusion: whenever a relation among basic objects can be represented by a functional relation among the measures of the objects, the function will obey an invariance condition. If all the measures are on ratio scales, the invariance condition takes the form of (10) or (11), but in general each combination of types of scales induces its own form of invariance condition. Thus an appropriate invariance condition is a logical consequence of more fundamental hypotheses, and the invariance principle is positive, not normative, when a basic empirical relation exists.

There are, of course, many cases in which the problem is not to represent a specified empirical relation but to define a theoretical relation among basic objects by first defining a relation among their measurements and then attributing a corresponding relation to the objects themselves. Such a relation need have no empirical counterpart.

A price index P_t, for instance, defined in terms of the measures $P_{1t}, \ldots, P_{2t}$ of n individual prices at time t, is supposed to permit us to compare 'the average price level' (APL) at times 1 and 2 by comparing the values of P_1 and P_2. We say that APL_1 is higher than APL_2 if and only if $P_1 > P_2$, or, more generally, that $APL_1, \ldots, APL_k$ are in relation S_A if and only if $P_1, \ldots, P_k$ are in relation S_P. The relation S_A is not a preexisting relation *represented* by S_P but is *defined* by it.

For another example, let A be a set of social states in the sense of Arrow (1951) and $u_i: A \rightarrow R$ be the ith person's utility function ($i = 1, \ldots, n$). A social welfare function $w: A \rightarrow R$ is defined by $w(a) = f[u_1(a), \ldots u_n(a)]$, where f: $R^n \rightarrow R$ is some function that obeys specified conditions. We say that state b is 'socially better' than state c if and only if $f[u_1(b), \ldots, u_n(b)] > f[u_1(c), \ldots, u_n(c)]$. This comparison between social states does not exist independently of the comparison between the values of f.

In such cases the invariance principle is normative and an appropriate invariance condition (Osborne 1976a) must be imposed. This condition is usually only one of a set of conditions that the relation among measures must meet. In some cases the full set of conditions cannot be met by any such relation. Examples occur in the theories of price indexes (Fisher 1911) and social welfare functions (osborne 1976b).

An Application

The implications of (11) depends on the qualitative structure of the part of $C_1 \times C_2$ that can be physically realized and that yields positive output of commodity 3. There are a number of interesting special cases, of which we can consider only one.

Suppose that for all $t > 0$ such that f is defined,

$$f(x_1, x_2) > 0 \text{ implies } f(tx_1, tx_2) > 0. \qquad (12)$$

(The restriction on t follows from the necessity that tc_k be in C_k when $c_k \in C_k$.) Then f is homothetic: for if (x_1, x_2) and (y_1, y_2) are on the same isoquant, so that $f(x_1, x_2) = f(y_1, y_2)$, then by (11) with $t_1 = t_2 = t$, $f(tx_1, tx_2) = f(ty_1, ty)$; so (tx_1, tx_2) and (ty_1, ty_2) are on the same isoquant.

Homotheticity follows from the invariance condition and the assumption stated in (12). This assumption can be recast into a weaker form that is common in theoretical work (let (x_1, x_2) and (y_1, y_2) lie on the lowest positive isoquant and confine t to values greater than 1). The invariance condition follows, under mild conditions, from our

schema. Essentially, then, empirical evidence against homotheticity would indicate defects in the schema. One obvious potential defect is the assumption of ratio scales, which, after all, still awaits a searching examination. The alternatives to ratio scales that have received much attention in measurement theory are ordinal, interval, and log-interval scales. Ordinal and interval scales are familiar in economics from utility and expected-utility theories. Log-interval scales h are unique up to transformations $h \rightarrow rh^s$, r, $s > 0$. The literature of economics would look very different if any of these alternatives were though appropriate for commodities. Suppose the measure of commodity 1 were unique up to the transformation $x_1 \rightarrow r_1 x_1^{s1}$, for instance. Then the price elasticity of demand, e, would be unique up to the transformation $e_1 \rightarrow s_1 e_1$, and the demand for commodity 1 would be more or less elastic than that for commodity 2 depending on our arbitrary choice of scales for measuring the commodities. This does not, of course, justify the assumption of ratio scales, but it indicates how radical the effects of abandoning the assumption would be.

Another potential defect in the schema is the assumption that a production function exists – or, in other words, that a production process can be modelled by a real-valued function of real variables. Georgescu-Roegen has energetically disputed this assumption on a number of occasions (e.g. 1970). But to follow up his line of thought would take us out of the realm of invariance into that of production.

See Also

- ▶ Homogeneous and Homothetic Functions
- ▶ Meaningfulness and Invariance

Bibliography

Aczel, J. 1966. *Lectures on functional equations and their applications*. New York: Academic.

Arrow, K.J. 1951. *Social choice and individual values*. New York: Wiley.

Bridgman, P.W. 1922. *Dimensional analysis*. New Haven: Yale University Press.

de Jong, F.J. 1967. *Dimensional analysis for economists*. Amsterdam: North-Holland Publishing Co.

Falmagne, J.C., and L. Narens. 1983. Scales and meaningfulness of quantitative laws. *Synthese* 55: 287–325.

Fisher, I. 1911. *The purchasing power of money*. New York: The MacMillan Co.

Georgescu-Roegen, N. 1970. The economics of production. *American Economic Review* 60: 1–9.

Jevons, W.S. 1965. *The theory of political economy*, 5th ed. New York: Augustus M. Kelley.

Krantz, D.H., R.D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of measurement*, vol. I. New York: Academic.

Kurth, R. 1965. A note on dimensional analysis. *American Mathematical Monthly* 72: 965–969.

Leontief, W. 1947. Introduction to the internal structure of functional relationships. *Econometrica* 15: 361–373.

Luce, R.D. 1978. Dimensionally invariant numerical laws correspond to meaningful qualitative relations. *Philosophy of Science* 45: 1–16.

Luce, R.D., and M. Cohen. 1983. Factorizable automorphisms in solvable conjoint structures, I. *Journal of Pure and Applied Algebra* 27: 225–261.

Luce, R.D., and J.W. Tukey. 1964. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology* 1: 1–27.

Osborne, D.K. 1976a. Unified theory of derived measurement. *Synthese* 33: 455–481.

Osborne, D.K. 1976b. Irrelevant alternatives and social welfare. *Econometrica* 44: 1001–1015.

Pfanzagl, J. 1968. *Theory of measurement*. New York: Wiley.

Popper, K. 1963. *Conjectures and refutations*. London: Routledge & Kegan Paul.

Ramsay, J.O. 1976. Algebraic representation in the physical and behavioral sciences. *Synthese* 33: 419–453.

Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton: Princeton University Press.

Whitney, H. 1968. The mathematics of physical quantities. *American Mathematical Monthly* 75: 115–138, 227–256.

Transition and Institutions

Peter Murrell

Abstract

This article examines the evolution of institutions and economists' thinking on institutions during transition. Early in transition,

institutions were virtually ignored in the majority of normative prescriptions, but were central in the evolutionary institutional approach. Later, after events influenced intellectual developments, institutions were at the centre of analysis. Growth is strongly related to institutional construction. Transition countries built institutions speedily but with marked variation across countries. Legal systems and independent governmental agencies were sources of institutional growth, while government bureaucracies and informal mechanisms detracted from institutional growth. In China, reforms addressed problems that institutions usually do, but in unusual ways.

Keywords

China; Economics in; Colonialism; Corruption; Creative destruction; Dual track liberalization; Evolutionary-institutional view; Institutional development; Institutions; Liberalization; Market institutions; Privatization; Resource curse; Rule of law; Shock therapy; Social capital; Stabilization; Transition

JEL Classification

P2; P3; N4; O17; K0

'Transition' is the widely accepted term for the thoroughgoing political and economic changes that followed the fall of communism in Eastern Europe (EE) and the Soviet Union. Some 29 countries are involved in this continuing process, which began in 1989–1991 and involves the types of transformations that usually took a century or more in today's developed countries. A related, but distinct, process has been under way in China since 1978.

Transition has been coterminous with a remarkable change in emphasis within economics. In 1989, to highlight the importance of institutions was a distinctly minority activity. Now, institutions are at the heart of both research and policy discussions. Similarly, while many early influential analyses of transition virtually ignored institutions, current discussions place them at the centre. Developments in transition countries made an important contribution to general trends within economics (Roland 2000).

This article focuses on institutions in the transition process and in economists' deliberations on that process. It begins with early normative prescriptions, in which institutions were virtually ignored by the majority of contributors, and then examines changing views on institutions, showing how events on the ground influenced intellectual developments. We then provide basic facts on institutional development, describing the impressive progress that has been made, which suggests modification of the standard assumption that institutional construction must be slow. Nevertheless, there is marked divergence across countries. This article examines the sources of institutional growth and the 'great divide' between the successful and the unsuccessful institution builders (Berglof and Bolton 2002). It closes by considering the seemingly anomalous case of China, showing that the anomaly is more apparent than real. China's reforms addressed the problems that institutions address everywhere, but in ways that are not recognizable to those using a first-best institutional template.

Ideas and Institutions in the Earliest Phase of Transition

Economists began to deal with the transition with a deluge of normative prescriptions. The majority view in its most stark incarnation came to be called shock therapy, the notion that the best way forward was as fast as possible on all fronts, taking advantage of a political window of opportunity. These types of reforms were certainly the aim of the first post-communist governments and their Western advisers in Poland, Russia and many other countries. According to shock therapy's proponents, the soon-to-be-observed dissonance between objectives and follow- through was variously due to the absence of a clear vision, lack of willpower, and a nefarious political opposition.

Institutions were ignored within the shock therapy approach for a variety of reasons. They could be built so easily that they did not require much attention (Sachs 1991). They were not deemed

important enough to mention (Blanchard et al. 1991). They would take so long to develop that other elements of policy came first (Fischer and Gelb 1991). Or, they could not be built without first creating the actors who would demand them within the political process (Boycko et al. 1995).

In shock therapy analysis, political economy considerations led to emphasis on the destruction of the old institutions and trumped any concerns about the dangers of an institutional void. Macroeconomics governed microeconomic institutional change, as exemplified by the International Monetary Fund's short-term focus on raising taxes in Russia, while largely ignoring sensible tax reforms (Black et al. 2000). Rapid liberalization was advocated, while downplaying its effects on the governance of contractual relations. The transaction costs of ownership change and corporate governance after privatization were deemed of secondary importance.

When economic performance in the early years of transition proved disappointing, diagnoses followed the earlier analyses: strong, but necessary, stabilization programmes had led to recessions (Blanchard et al. 1994); that is, in the early 1990s, the most influential analyses did not associate the steep, sometimes catastrophic, recessions with institutional problems. For example, such analyses led to the conclusion that liberalization, privatization and stabilization should move even faster in Russia in 1992 than they had in Eastern Europe 2 years earlier.

Kogut and Spicer (2004; 2005) use numerical citation analyses to analyse patterns in the early economics literature on transition. They document the links among a core group of economists subscribing to the shock therapy approach. This group had strong connections to the international financial institutions and the US government, and were able to spread their views in reforming countries under the auspices of these powerful organizations. Kogut and Spicer also identify dissenters from this point of view, in particular Murrell (1992), Dewatripont and Roland (1992), and McKinnon (1991). Early in transition the dissenting view was labelled evolutionary or gradualist, but was later given the much more felicitous name, evolutionary-institutional (Roland 2000).

The evolutionary-institutional view emphasized the importance of institutions, suggesting, for example, that the nature and timing of liberalization, privatization and stabilization depended critically on the existing institutional framework. Some institutions were prerequisites for a functioning market economy, and the absence of these might necessitate the slowing of reforms. Because new market economy institutions were hard to create, it might be better to use crude second-best institutions, even some of the old ones, while maintaining a focus on building new ones. This might lead to a two-sector approach, where a nascent private sector was governed by new institutions, while some of the old mechanisms of governance prevented convulsions in the old state sector, negatively affecting the development of the new capitalism. This approach was particularly congenial for those who thought that the growth of the new private sector was crucial (Kornai 1990) or whose advice reflected elements of Chinese reforms (McMillan and Naughton 1992).

The suggestion that economic reform should be gradual was a conclusion, rather than a starting point. It grew out of analyses that were standard in the literature (North 1990). Because ideas and organizations adapt to an institutional framework, there is no certainty of an immediate functional response to new institutions. Difficulties in creating new institutions suggest a wariness of quick reforms when their success depends on functioning institutions. Instead, a nascent private sector produces the most nimble response in a new environment of fast-changing institutions.

Evolving Ideas on the Role of Institutions

Events changed ideas. All transition countries experienced deep recession. Recovery began after several years, with its inception unrelated to any specific policy initiatives. If anything, recovery began on retreat from the earlier policies. The degree of adherence to standard policy advice could not explain the cross-country pattern of recession and growth.

Although these facts were consistent with the evolutionary-institutional view, the most influential contribution in changing the terms of debate was a paper coauthored by one of shock therapy's main proponents (Blanchard and Kremer 1997). Undoubtedly, this paper had such a large effect because one of its authors was an influential economist who had previously attributed little importance to institutions (see Blanchard et al. 1991; Blanchard et al. 1994; and the review of the latter in Murrell 1995). The paper formalizes ideas already present in the earlier evolutionary-institutional literature in a simple, but powerful model. The model highlights the incentives to break agreements in the absence of effective governance, leading to a loss of production. Output decline comes later but is larger where the complexity of old production relations is greater. If opportunities improve over time, the model generates a U-shaped path for production. These predictions are consistent with the comparative profiles of growth in the transition countries, with recession initially steeper but ultimately shallower in Eastern Europe than in the former Soviet Union (FSU).

There is much to learn about the relationship between institutional change and production decline, but there is general agreement in some areas. Pre-transition institutions contributed to enterprise productivity. These institutions offered credibility in the negotiating of agreements, contract enforcement, specification of control rights over assets, mechanisms for the generation and allocation of working and investment capital, and many other services. When the communist systems fell apart and market institutions were still on the drawing board, these crucial services were no longer supplied. The lack of institutional support was particularly critical at the beginning of transition for several reasons: socialist firms were large, implying a need for sophisticated governance mechanisms; inter-firm relationships were highly particularized, implying great potential for hold-up problems; and necessary adjustments were enormous, implying the need for effective financial markets.

Even without effective institutions, production rebounded due to the spontaneous growth of private sector opportunities. Nascent small businesses could take advantage of these opportunities if they received a minimal amount of institutional support, that is, protection from extreme criminality, prevention of civil chaos, and the benign neglect of the state. Such businesses develop their own self- enforcing agreements and do not need sophisticated courts or contract law. Physical possession solves many concerns about property rights. Closely held firms that are self-financed do not need corporate governance institutions.

But to rebound from recession is not the same as sustained growth. The latter requires more than the benign neglect of the state: it requires a set of institutions that support non-self-enforcing agreements, secure property without possession, enable firms to expand beyond the limits of self-finance, and undertake many other activities that are not feasible without effective rules of the game. While such ideas seem commonplace now, they were not to the fore in the debates at the start of transition, except in the evolutionary-institutional perspective.

In addition to the institutional interpretation of the causes of collapse and recovery, two further factors contributed to economists' changing views. First, econometric studies showed that differences in the application of the standard policies did not explain differences in economic performance (for example, de Melo et al. 2001; Falcetti et al. 2002). Second, variations in performance became more noticeable in the trajectories out of recession. Countries appeared to be sorting themselves into two groups. Those in EE were generally performing better than those in the FSU, but there were enough exceptions (for example, the Baltics, Serbia) to suggest that the EE–FSU distinction was not the key. As Berglof and Bolton (2002, p. 77) noted, 'A growing and deepening divide has opened up between transition countries where economic development has taken off and those caught in a vicious cycle of institutional backwardness and macroeconomic instability'.

Beck and Laeven (2005) were the first to test this new institutional paradigm of growth in transition in a rigorous framework, although their study is naturally characterized by a paucity of data points. They find that there is very large

divergence in the performance of transition countries and that institutional development is the key factor in explaining the divergence. Moving from Russia's level of institutional development in 1996 to Poland's level would lead to a growth rate increase of 4.4% a year. In contrast, differences in policies are unimportant. Papers studying privatization, agricultural markets and foreign direct investment contain results on economic performance at a more disaggregated level that complement those of Beck and Laeven. While relative neglect of institutions characterized the early stages of transition, the centrality of institutions is now conventional wisdom.

Evolving Institutions

One reason why institutions were not emphasized in early transition was the widely held assumption that institutional construction would be very slow. The transition countries provided an ideal testing ground for this assumption. Having rejected a set of old institutions and turned to creating new ones, how fast and how successful could institutional construction be? The answer, for some countries only, is: surprisingly quick and successful. Transition experience refutes one element of conventional wisdom, that institutional development is inevitably very slow, while bolstering another, that failure is commonplace.

Murrell (2003) concluded that there had been widespread, large, continuing improvements in institutional quality from 1990 to 2000. An updating especially for this article extends this analysis to 2004, using the popular institutional measures developed by Kaufmann et al. (2005). This updating shows that institutional scores for transition countries as a whole are no better and no worse than one would expect given levels of economic development. This is remarkable, since it implies that in less than 15 years the transition countries built institutions that match those in countries that have had capitalist systems for much longer. For example, on the rule of law, Hungary, Slovenia and Estonia are comparable to Chile, Israel, Greece, Italy, Spain and Taiwan. On regulatory quality, Estonia ranks above Sweden, while Hungary, Lithuania, Slovakia, Latvia and the Czech Republic are grouped with the United States, Japan, Italy and Spain.

These results, which are based on expert opinions and surveys, are supported by studies examining the micro details of institutional development. Djankov et al. (2002) collect data on highly specific aspects of the functioning of legal systems, such as collecting on a bad cheque. They find that the ex-socialist countries fare better than both French-legal-origin and German-legal-origin countries. Pistor et al. (2000) examine the quality of laws on shareholder and creditor rights, finding the transition countries superior to many developed economies.

The second distinctive feature in institutional development is the divergence between one group of countries whose institutions are at a comparatively high level and improving and another group that has not crossed the great divide and is even losing some of the gains from the 1990s. By 2004, the EE–Baltic group has institutional scores higher than expected, given general levels of economic development on all six of the Kaufmann et al. (2005) indicators, and these scores improved dramatically in the preceding decade. The Commonwealth of Independent States (FSU minus the Baltics) scores below expected levels and has been regressing from 1996 to 2004, after showing remarkable signs of institutional improvement in the early 1990s.

It is difficult to exaggerate the importance of this empirical evidence on basic hypotheses on institutional development. Before transition, the assumption was that modern institutional development is a very long process, fraught with the possibility of failure. The first element of this assumption has been refuted. In the years before 1990, capitalism and democracy were absent in EE and the USSR. Then there was a mammoth fall in national income, due to institutional lacunae. Yet now a large group of countries seems set on the road to sustainable institutional development. In contrast, the second element of standard assumptions has been verified. In a significant number of transition countries, slow initial progress on institutional development has been followed by severe regression.

The Sources of Institutional Development

There are two alternative perspectives to take when viewing the sources of institutional development. First, one can analyse which country-level factors best explain aggregate institutional outcomes. Second, one can ask which particular mechanisms or organizations inside a country contributed most to institutional performance. Evidence on both is only currently being generated, and is very scant. This is true both of transition and in general.

Beck and Laeven (2005) have carried out the most systematic study of the causes of aggregate institutional development in transition. They find two principal determinants: the strength of the incumbent socialist elite and the importance of natural resources (the resource curse). Both are negatively related to improvement in institutions. They also confirm the analysis of Black et al. (2000) that certain types of privatization might have been inimical to institutional development. Early macroeconomic policies, belonging to the FSU and being eligible for the European Union do not affect institution building. These negative results are important since they reject prominent hypotheses. One popular theory not explored by Beck and Laeven is that colonial heritage might have influenced institutional development, particularly in the case of countries influenced by the Austrian, Ottoman or Russian empires.

One can also ask which particular mechanisms contributed most to institutional performance (Murrell 2003). Formal institutions have played a more beneficial role than informal institutions, such as culture or pertinent elements of social capital. Of the formal institutions, political and legal structures and independent governmental agencies contributed relatively more to institutional development. State administrative bodies detracted from institutional performance, changing slowly and contributing to relatively high levels of corruption. These facts are generally consistent with the old Schumpeterian message of creative destruction, but applied to non-market organizations.

One very surprising feature of transition is the relatively strong role of some legal institutions. A series of empirical observations on the courts suggests a divergence from prevailing views on the role of the legal system (see the essays in Murrell 2001 for example). The legal system has never been identified as playing a strong role in developing countries, and transition was not conducive to the effectiveness of the law. Yet current evidence suggests that it is easier than usually assumed to fashion a legal system that facilitates economic processes, even when that system is far from the standards of developed countries.

Institutions in the Chinese Reforms

On the surface, Chinese reforms might be cast as a refutation of the above. China began its reforms with a basic constraint on institutional change – movement from the existing system could not be too great or too fast. This meant that new institutions would not be best practice, but had to be incremental variations on existing ones. Hence, China does not fare well when matched against standard criteria for judging institutions. This stands in contrast to the astounding success of the Chinese economy.

Nevertheless, China's reforms can be interpreted as bolstering the basic conclusion of the centrality of institutions in transition. China created successful, transitional institutions (Qian 2003). By experiment, by confining itself to incremental changes that could be easily understood, and by implementing Pareto- improving changes in the early years of reform, China pursued a deft, but previously untrodden path of institutional change.

Qian (2003) provides examples of these transitional institutions. China implemented a dual-track approach to liberalization, which led to markets in above-plan production, but kept quotas and controlled prices on the levels of production that had existed before reforms. This promoted efficiency at the margin, while endorsing the existing set of informal rights to infra-marginal production, thus protecting the welfare of those who otherwise might have lost heavily from reforms. A highly distinctive ownership form appeared, township and village enterprises (TVEs), which

played a significant role in China's growth in the first two decades of reform. TVEs can be interpreted as a mechanism for protecting decentralized property rights when the state is unable to guarantee more formal ones for private owners. Anonymous banking served as a commitment device, limiting government predation by reducing information flows. This arrangement can be understood as a crude substitute for the protection of financial property rights when the independence of the legal system is not a real possibility in the short-run.

Therefore, China constructed mechanisms to address the problems that institutions address in successful countries. However, those mechanisms would not look familiar when matched against best practice in developed countries. As the evolutionary-institutional perspective emphasized, it is fruitless to try to imitate best practices when human capital and institutional capability are not sufficient. In such situations, it might be best to deploy a set of transitional institutions, much more suited to the particular circumstances of a country and its capabilities. This observation resonates with the experience of EE and the FSU reviewed above. Countries with less benign starting points for creating best-practice institutions were doomed to fail in the process, while others could succeed given the right political and human capital preconditions.

Of course, there was a reason why in early transition best-practice institutions were advocated for all countries. It was feared that a country might find it hard to replace transitional institutions once they were set in place, becoming trapped at a low level of development. Whether this fear was ultimately justified will be addressed by Chinese experience in the coming decades.

See Also

- ▶ Command Economy
- ▶ Dual Track Liberalization
- ▶ Great Divide
- ▶ Market Institutions
- ▶ New Institutional Economics
- ▶ Output Fall – Transformational Recession

Bibliography

Beck, T., and L. Laeven. 2005. *Institution building and growth in transition economies*, Policy Research Working Paper 3657. Washington, DC: World Bank.

Berglof, E., and P. Bolton. 2002. The great divide and beyond: Financial architecture in transition. *Journal of Economic Perspectives* 16(1): 77–100.

Black, B., R. Kraakman, and A. Tarassova. 2000. Russian privatization and corporate governance: What went wrong? *Stanford Law Review* 52: 1731–1808.

Blanchard, O., and M. Kremer. 1997. Disorganization. *Quarterly Journal of Economics* 112: 1091–1126.

Blanchard, O., R. Dornbusch, P. Krugman, R. Layard, and L. Summers. 1991. *Reform in Eastern Europe*. Cambridge, MA: MIT Press.

Blanchard, O., K.A. Froot, and J.D. Sachs. 1994. *The transition in Eastern Europe*. Chicago/London: University of Chicago Press.

Boycko, M., A. Shleifer, and R. Vishny. 1995. *Privatizing Russia*. Cambridge, MA: MIT Press.

de Melo, M., C. Denizer, A. Gelb, and S. Tenev. 2001. Circumstances and choice: The role of initial conditions and policies in transition economies. *World Bank Economic Review* 15: 1–31.

Dewatripont, M., and G. Roland. 1992. The virtues of gradualism and legitimacy in the transition to a market economy. *Economic Journal* 102: 291–300.

Djankov, S., R.L. Porta, F. Lopez-de-Silanes, and A. Shleifer. 2002. *Courts: The Lex Mundi project*, Working Paper No. 8890. Cambridge, MA: NBER.

Falcetti, E., M. Raiser, and P. Sanfey. 2002. Defying the odds: Initial conditions, reforms, and growth in the first decade of transition. *Journal of Comparative Economics* 30: 229–250.

Fischer, S., and A. Gelb. 1991. The process of socialist economic transformation. *Journal of Economic Perspectives* 5(4): 91–105.

Kaufmann, D., A. Kraay, and M. Mastruzzi. 2005. *Governance matters IV: Governance indicators for 1996–2004*. Washington, DC: World Bank.

Kogut, B., and A. Spicer. 2004. Critical and alternative perspectives on international assistance to post-communist countries: a review and analysis. The World Bank, Operations Evaluation Department. Online Available at http://www.worldbank.org/oed/transitioneconomies/docs/literature_review.pdf. Accessed 31 May 2007.

Kogut, B., and A. Spicer. 2005. *Taking account of accountability: Academics, transition economics, and Russia*. Paris: INSEAD.

Kornai, J. 1990. *The road to a free economy*. New York: Norton.

McKinnon, R. 1991. *The order of economic liberalization: Financial control in the transition to a market economy*. Baltimore, MD: Johns Hopkins University Press.

McMillan, J., and B. Naughton. 1992. How to reform a planned economy: Lessons from China. *Oxford Review of Economic Policy* 8: 130–143.

Murrell, P. 1992. Evolution in economics and in the economic reform of the centrally planned economies. In *Emerging market economies in Eastern Europe*, ed. C.C. Clague and G. Rausser. Cambridge, MA: Basil Blackwell.
Murrell, P. 1995. The transition according to Cambridge, Mass. *Journal of Economic Literature* 33: 164–178.
Murrell, P. 2001. *Assessing the value of law in transition economies*. Ann Arbor: University of Michigan Press.
Murrell, P. 2003. The relative levels and the character of institutional development in transition economies. In *Political economy of transition and development: Institutions, politics and policies*, ed. N. Campos and J. Fidrmuc. Boston/Dordrecht/London: Kluwer.
North, D. 1990. *Institutions, institutional change, and economic performance*. Cambridge: Cambridge University Press.
Pistor, K., M. Raiser, and S. Gelfer. 2000. Law and finance in transition economies. *Economics of Transition* 8: 325–368.
Qian, Y. 2003. How reform worked in China. In *In search of prosperity: Analytic narratives on economic growth*, ed. D. Rodrik. Princeton: Princeton University Press.
Roland, G. 2000. *Transition and economics: Politics, markets, and firms*. Cambridge, MA: MIT Press.
Sachs, J. 1991. Poland and eastern Europe: What is to be done? In *Foreign economic liberalization: Transformation in socialist and market economies*, ed. A. Koves and P. Marer. Boulder: Westview Press.

Transitional Labour Markets: Theoretical Foundations and Policy Strategies

Günther Schmid

Abstract

In a normative perspective, the theory of transitional labour markets (TLM) considers the labour market as a social institution supporting and ensuring 'full employment' not only in terms of income security, particularly in times of unemployment (freedom from want), but also in terms of the capability to freely choose and to develop a career over the life course. This capability should equally hold for men and women, and it should also include unpaid but socially highly valued phases of work (freedom to act). TLM theory also argues for a new analytical approach to study the dynamics of labour markets: instead of concentrating on stocks (e.g. employment or unemployment rates), flows should come to the fore, in particular transitions from one employment status to the other, including combinations of work and education or work and unpaid care. This entry elaborates on these analytical and normative foundations of TLM theory and develops respective policy strategies to manage social risks related to the main critical life course transitions.

Keywords

Transitional labour markets; Non-standard employment; Flexibility; Security; Labour market policy; Social policy; Europe

JEL Classifications

J21; J38; J4; J48; J68; R28

Introduction

The theory of transitional labour markets (TLM) emerged in the mid-1990s as a response to the European unemployment crisis, which then was considered by the majority of researchers as a reflection of 'Eurosclerosis'. This view led to the corresponding recommendation of deregulation as a solution, promoted in particular by the OECD. Although partly concurring with the diagnosis of a fundamental deficit in flexible employment relationships, the TLM concept contradicted the therapy of deregulation by arguing for a complete reorientation of labour market institutions towards a new concept of 'full employment'. Whereas the old concept referred to a situation of full-time and long-term jobs for all members of the working age population, implicitly however thought only for male breadwinners, TLM's new concept of 'full employment' explicitly included all women and extended the 'full employment' definition by a vision of flexible employment relationships over the life course (Schmid 1995).

In a normative perspective, TLM considers the labour market as a social institution supporting and ensuring 'full employment' not only in terms of income security, particularly in times of unemployment (freedom from want), but also in terms of the capability in freely choosing career perspectives over the life course that include unpaid but socially highly valued phases of work (freedom to act). Apart from this normative perspective, TLM theory also argues for a new analytical approach to study the dynamics of labour markets: instead of concentrating on stocks (e.g. employment or unemployment rates), flows should come to the fore, in particular transitions from one employment status to the other, including combinations of work and education or work and unpaid care. The policy response of TLM theory is the concept of social risk management as a pathway to prevent, mitigate or to cope with the social risks related to life course transitions (Schmid 2002; Schmid and Gazier 2002).

This entry examines the analytical and normative foundations of TLM theory and explains policy strategies to manage social risks related to the main critical life course transitions. TLM's history, methodology and empirical work have previously been reviewed (see: van den Berg and de Gier 2008; Gazier and Gautié 2009; Rogowski 2008).

The Analytical Basis of Transitional Labour Markets

The theory of TLM aims at giving – in the new world of work – labour market policy in general and the new European employment strategy in particular a clear analytical and normative framework. According to this view, the labour market has to be considered as a system of employment transitions over the life course: transitions (flows) between education and employment or 'school-to-work transitions'; transitions within employment or 'job-to-job transitions'; transitions between employment and unemployment; and transitions between the two other forms of economic inactivity, unpaid household-work or the status of being severely disabled, ill or in retirement. In this perspective, stocks have to be considered as product of flows into or out of un/employment or inactivity and the duration of being in the respective status.

The TLM perspective breaks with the illusion that all or at least most unemployed come immediately from the status of employment, or – vice versa – that most unemployed transit sooner or later back into employment (Schmid 2008, pp. 165–212). Flow, rather than stock statistics (e.g. Eurostat 2016) are helpful – for instance – in distinguishing between structural and cyclical un/employment: if unemployment stagnates because the substantial outflows from unemployment to employment are counterbalanced by inflows from inactivity to unemployment, the development can be judged as a recovery on the labour market despite the stagnating unemployment levels because large numbers of inactive individuals have decided to take up the search for employment, thus being counted as unemployed. The decision to search for work might indicate that these individuals earlier regarded their chances on the labour market as so low that they did not even search for work (discouraged workers). The large outflows from unemployment to employment indicate that the demand for labour has indeed increased. The situation is different if unemployment is stagnating because there are only minimal outflows from unemployment. In that case, flow statistics indicate that the labour market is not improving.

The OECD, looking at the institutional determinants of transitions, identified two interesting results from the TLM point of view: first, large gross jobs and worker flows partially reflect better job opportunities available to workers due to an enhanced job-matching process. Voluntary job changes tend to be connected with wage premiums whereas workers facing involuntary separations suffer from wage penalties in case of reallocation, even if not connected with a spell of unemployment. Second, unemployment benefit generosity appears to have a positive impact on average gross worker flows. A ten-percentage-point increase in the average net benefit replacement rate would increase gross worker

reallocation by about one percentage point (OECD 2010, pp. 169–170).

From a TLM point of view, such flow statistics, however, are still insufficient. In one of the first summary papers on TLM, Gazier and Gautié (2009, p. 3) bring the real need of information to the point: '[. . .] the boundaries between the (supposedly) stable departure and arrival positions may become blurred. In particular, the traditional distinction between three basic positions or "states" (employment, unemployment, and inactivity) is weakened by the development of a complex set of intermediate positions, depending on public schemes and programmes such as short-time working, progressive early retirement or subsidized employment. The deliberate and concerted management of these intermediary positions is [...] at the heart of the TLM approach. One has to remark that beyond these intermediate states, the very frontiers of firms also become blurred [..].'

Of special importance for policy considerations are transition studies targeted towards 'good' and 'bad' transitions, where 'good' and 'bad' can be defined by various quality criteria: wages (low vs. high pay), working time (part-time vs. full-time) and labour contract (temporary vs. open-ended or dependent vs. self-employment). Furthermore, these criteria themselves have to be put into context, especially in terms of career perspectives such as 'bridges' vs. 'traps', 'stepping stones' vs. 'dead-end jobs', 'integrative' or 'maintenance' or 'exclusionary' transitions in the life course perspective (O'Reilly 2003, pp. 1–48). Education and training are also of critical importance for successful transitions over the life course. In order to respond to a lack of employment opportunities and to risks of unemployment and social exclusion, individuals need to be enabled to organise and pass through multiple transitions between working and learning throughout their life course (Schömann and O'Connell 2002, p. 5). Transitions over the life course between various employment statuses (permanent, temporary, self-employment, non--employment, education or training) are also central from the TLM point of view (e.g. Calandrino and Gagliarducci 2004). Furthermore, the gender impact of transition dynamics is – according to TLM's new full employment goal – of special importance and a subject of many studies (e.g. Guergoat-Larivière and Erhel 2010; Leschke 2015).

For the design of labour market policy a crucial issue is whether placement services for unemployed or inactive people should concentrate more on jobs than on long-term employability. Many studies confirm that placement strategies according to the 'work first' principle lead to sub-optimal results. In principle 'work first' might be a meaningful orientation, especially towards the low skilled for whom training *on* the job seems to be more effective than training off the job. Efficiency-oriented employment services, however, have to look beyond a quick placement and towards ensuring sustainable employability and employment security with a high productivity potential, with the latter requiring continuous education and vocational training over the life course, in particular for low and medium skilled workers (Schmid 2014).

The focus on positive transitions over the life course for low skilled workers is also justified by repeated observations that it is the low skilled who are disproportionately affected by increasing forms of non-standard employment (NSE), including part-time work, temporary work (fixed or project-based contracts, casual labour, minijobs or even zero-hour contracts), triangular employment relationships through temporary agencies or subcontracting companies and self-employment. In EU-28, the NSE rate increased from about 21.5% (1998) to 25.8% (2014); in other words, about a quarter of the working-age population works – controlled for overlaps – either in different forms of part-time, temporary work or self-employment. Whereas temporary work and self-employment rates have stagnated or even decreased since the recession of 2008/09, part-time employment is still on the rise in almost all EU member states, driven mainly by the service and knowledge economy (Schmid 2016; Schmid and Wagner 2017).

Empirical studies about transition dynamics and their determinants, such as the volumes by Anxo et al. (2007, 2008), Lassnigg et al. (2007) and Muffels (2008), have investigated how national

regulatory and social protection systems promote and support transitions that are life course oriented, facilitate a better work-life balance of individuals and households and strengthen the social cohesion of European societies. Transition patterns within work and between various employment relationships are explained against the backdrop of a rapidly changing societal context due to ageing, individualisation and globalisation, highlighting the emergence of non-standard forms of employment, and in particular the supposed 'scarring' effects of these new employment forms on the future career and their alleged 'stepping stone' functions in 'standard' jobs. Muffels (2008, p. 390) addresses the need of fundamental institutional reforms that would be able to deal with the changing world of work and argues in particular for reconstructing social cohesion on the basis of 'competitive solidarity'. This conclusion leaves open the normative issues of which direction and with what kinds of new institutional arrangements such solidarity might be established.

The Normative Basis of Transitional Labour Markets

The normative principles of TLM draw on theories of justice (Rawls 2001; Dworkin 2000), social choice (Sen 2009) and behavioural economics (Kahneman 2011), as set out by Schmid (2008, pp. 224–239). The core idea is to *empower individuals* to take more risk during the life course: first, by making not only work pay but also by *making transitions pay* through extending the social insurance principle beyond unemployment and including volatile income risks connected with other critical events over the life course; second, by making not only workers fit for the market but also by *making the market fit for workers* (slogans originally coined by Gazier) by enhancing employers' and employees' capacity to adjust to uncertainties by investing in individual competences as well as in the workplace environment (Auer and Gazier 2006; Gazier 2003; Jørgensen and Madsen 2007; Schmid 2015).

In following Sen's conclusion that it is not resources but capabilities, i.e. the shift from means of living to actual opportunities (Sen 2009, p. 253), TLM suggests a new basis for active labour market policy (ALMP) emphasising active *securities*, giving people hand-ups instead of only hand-outs. 'Active' means first, investing in people versus passive charity, as in pure market economies; and second, protecting people's investments versus protecting jobs, as in pure socialist economies. From this perspective a generous income replacement for finding a new job is a productive investment (Acemoglu and Shimer 2000). Subsuming unemployment benefits – if properly designed – under passive labour market policy is a serious mistake because 'passive' connotes only the costs and not the benefits.

The second emphasis of TLM is its *life course orientation*. Its concept of '*careers*' acknowledges the right of an individual to a *development* perspective in contrast to the neoliberal concept of *'workfare'* that restricts work to an obligation in order to deserve transfers in the case of need. The right to a career also entails a voice in choosing jobs and working conditions in contrast to directing people into jobs in pure socialist economies. Modern labour market services, therefore, have to support *transition securities* beyond the employment-unemployment transition (Anxo et al. 2007; Schmid 2014; Zimmermann 2011).

This leads to a third emphasis, namely *to empower individuals* to change from one work-situation to another according to changes in the economy as well as according to individuals' changing preferences or work capacities over the life course. *Citizens should therefore have the right to transitions*. *'Work'*, in this context, includes all activities of social obligatory character, whether paid or not. Even participation in collective decision making should be considered as work because exercising voice in work-related decisions is an essential part of economic democracy (Hirschman 1970). An early example of a work-related right to exercise voice was the granting of time off to representatives of works councils. Other examples are the right to negotiated leaves of absence for training and sabbaticals, and the right to family-related furloughs such as parental or other care leave (Supiot 2001/2016).

Most people accept changes more easily if the risks are shared justly, which is why the theory of justice plays an important role in the TLM concept. Four principles of justice build the *normative pillars of risk-sharing*: first, *justice as fairness*, notably in access to jobs, with inequality only justified if the lot of the most disadvantaged improves; second, *justice as solidarity*, which means sharing responsibilities according to the type of risks and individual capacities; third, *justice as agency*, which means developing individual *and* institutional capabilities to enhance individual and regional autonomy, in other words, freedom to act; fourth, *justice as inclusion*, which means enlarging risk-sharing communities according to the interdependencies of economic and social life.

This normative backdrop of TLM forces researchers and policy makers to concentrate on *risky events* over the life course and to look at whether job-to-job transitions lead to *social integration*, *career development* or *social exclusion*. This requires, as already mentioned, analytical and empirical instruments to study transitions and multi-year transition-sequences or trajectories (Brzinsky-Fay 2011), to utilise intelligent transition matrices and to control individual transition sequences through proper statistical methodologies. The perspective of maximising employment opportunities draws the attention also to the adjustment potential of transitions *within* stable employment relationships (for instance, the transition from full-time work to part-time work or the combination of part-time work with part-time education or training and care for children or other dependents), which is of great – and often neglected – importance. Such *internal flexibility* has to be considered to be the functional equivalent to *external flexibility*, relevant particularly in view of the fact that most people still strive towards tenured positions. Evidence of stable if not increasing job tenures (Auer et al. 2005) does not contradict the TLM perspective.

This view also justifies a redefinition of full-employment targets. For example, the EU's Europe-2020 goal of raising the employment rate to 75% for the working-age population (20–64) stems from the demographic challenge of a declining labour force. TLM's main motivation for raising the full (nominal) employment target, however, is to allow individuals *and* employers greater variation in effective employment by providing active securities beyond the risk of unemployment, in particular through working time variation over the life course. The aim is a win-win strategy in which employers gain greater capacity of work-place adjustment and the workers greater autonomy in the choice of working time and employment.

Policy Strategies in Managing Labour Market Transitions and Risks

The life course approach of TLM's analytical framework leads to a focus on the main social risks that occur to all people over the life course (Schmid 2008, pp. 281–328): lack of earnings capacity (I), earnings insecurity (II), total loss of earnings (III), restricted earnings capacity (IV), and reduced earnings capacity (V). 'Restricted' earnings capacity presupposes a potential full capacity to enter gainful (full-time) employment, but that capacity is intermediately restricted by social obligations of (mostly) unpaid work; 'reduced' earnings capacity is always an individual lack of 'full' capacity due to (partial) disability, illness or physical exhaustion. Dealing with these risks requires distinctive policies for different sorts of transitions.

Managing School-to-Work Transitions

School-to-work transitions largely determine the range of employment opportunities over the life course and have been extensively studied, especially in response to pervasive youth unemployment in Europe (e.g. Brzinsky-Fay 2011); the STYLE research network, a recent example, coordinated by Jackie O'Reilly also deploys a TLM approach (http://www.style-research.eu/).

Managing the basic risk in the transition phase from school-to-work, i.e. the *lack of earnings capacities* or even the risk of *un-employability*, means – first of all – sufficient *inclusive* investments on 'human *and* social capital', especially in general competences like reading and

mathematical skills, communication skills, learning abilities and secondary virtues like endurance and ambiguity tolerance. By emphasising human *and* social capital to develop earnings capacities, young people become fit for the market, able to raise their voice and to shape the market. A high earnings capacity is the best insurance against all other social risks that occur during the life course. Second, managing the risk of un-employability over the life course cannot only consist of providing high levels of formal education. Education has to be related to market needs, i.e. to the skills required to produce the goods or services consumers demand.

The TLM concept suggests four elements to ensure employability: the combination of learning, working, earning *and* identity building; the combination of job-specific *and* general skills; the reduction of information asymmetries by voice *and* trust; and a fair risk sharing of costs and benefits related to the investment.

New jobs often require new skills (European Commission 2012). But it would be a mistake to think that all these new skills require tertiary education at universities: Time served in formal education is not enough; what counts is what you can do with what you know. This becomes all the more true with the Internet revolution and its rapid access to all the passive knowledge that one may need. Furthermore, skills acquired in the formal education system may not suffice over the whole life course. With lifelong learning there is more at stake than a further extension of formal schooling, particularly in view of complementarities in the learning processes (Heckman 2008). It is also wrong to believe that all young people prefer 'knowledge'-related work. Many prefer practical work and work with which they can connect some meaning and which gives them a personal identity.

This is why it makes sense to establish *dual learning systems* in particular – but not necessarily only – in the transition phase from school to work. Dual learning systems are the paradigm for the ideas of TLM, offering *institutional bridges* between labour market work and work or life outside the labour market to ensure a *'work-life balance'*. Part of the underlying theory is the insight that human *and* social capital are not only built in schools but also on the job. The flipside of this insight is that the longer people remain jobless the more their acquired human *and* social capital deteriorates. So, everything has to be done to avoid or to reduce unemployment not only for youth but also for adults, including mature-aged workers.

Another essential element of this theory is *fair risk sharing* of the costs and benefits, here related to human capital investments but also relevant to other investments on employability, mobility or work-place adjustments. Employers will be reluctant to invest in training or education if people run away after the costly investment, or if these skilled people are then poached by employers that have not contributed to the investment. Employees will not invest in firm specific skills if their investment is not rewarded by fair wages, good working conditions and some job security. The state or the social partners (trade unions and employer associations) can play a crucial role in solving these conflicts by co-financing (in particular education infrastructure), by defining and controlling marketable quality standards, by wage coordination and by reasonable employment protection. Through the standardisation of training contents, social partners and governments can ensure high-quality standards through control and certification. Participation in the definition of quality standards by employers and employees (usually via their sectoral or occupational interest representations) and their effective control guarantee, on the one hand, that workers can trust that their skills are valued on the market, and on the other hand, that employers can rely on the competences of graduates entering the labour market. Certificated (i.e. legally acknowledged) skills are also 'marketable' and thereby enhance the mobility of workers.

European economies confirm the value of such dual learning systems. Youth unemployment is lowest in European countries with dual learning systems that connect their education system closer to the labour market. These countries are Austria, Denmark, Germany, Netherlands and Switzerland (Ebner 2012; Eichhorst et al. 2012); these countries also have a low incidence of youths 'Not in Education, Employment or Training' (NEET). On

average, increasing the share of upper secondary students that attend dual learning systems by one percentage point decreases NEET rates by about 0.04–0.09 percentage points (Eurofound 2012).

There are both demographic and technological reasons for boosting continuous education and training, and behavioural justifications for a universal lifelong learning insurance. The first behavioural argument is *uncertainty of returns for workers*. Our cognitive map values what we already possess much higher than what we might expect from risky investments (Kahneman 2011, pp. 289–99). This endowment effect is more relevant for low-skilled and low-income earners than for the high-skilled. The resulting risk aversion can only be overcome by extending the expectation horizon through conditional job security or through employment security; legal rights to lifelong learning and possibly legally guaranteed minimum levels of education would also extend the expectation horizon; finally the certification of acquired new skills belong to this solution which, ideally, would have to be put into the context of learning modules leading to a promising career perspective. For *employers*, investing in their low-skilled workers *entails the high risk that the returns of their investments might be zero or small* due to low learning capacities.

Finally, there are *information uncertainties* affecting both labour supply and labour demand. Workers are often faced with complete intransparency of the training market, while employers do not know what kind of skills they should invest in. More importantly, these uncertainties beget another uncertainty: The players of the lifelong learning game – workers, employers *and* the state – do not know beforehand where gains are going to accrue and where losses must be incurred. This observation holds true at micro- and macro-levels alike. The veil of ignorance, the insurance situation, is a given.

The solution to these information uncertainties can only be in learning by monitoring (Sabel 1994) through establishing *learning communities* at the local or regional level, in which all relevant actors – schools, training institutions, employers, and social partners – get involved. This involvement, however, needs to be organised in a form that makes actors committed and responsible (Korver and Oeij 2008; Schmid 2011, pp. 105–12).

Covenants are an established form of such negotiated flexibility and security, well-known under this term in the Netherlands. Best practice in continuing education and training is not common knowledge yet, but it probably already exists de facto, for instance, in Denmark (Lassen et al. 2006), and it may be the secret of successful local or regional labour market pacts or local strategic partnerships (Burroni et al. 2010). It is also likely to evolve, for the urgency of this overarching common goal at all levels of governance is pressing, e.g. in the EU policy initiative on *'New skills and jobs'* (European Commission 2012). In covenants, partners retain an exit option if the risk-taking appears excessive. Because the balance of costs and benefits might change at each step, the partners involved in the game must trust in the possibility that corrective measures can be taken in pursuit of the common goal through a fair process of re-negotiations (Korver and Schmid 2012).

Managing Job-to-Job or Within-Job Transitions

This leads us to the next main life course risk: *Earnings insecurity* due to volatile wages (related, for instance, to temporary jobs or temporary layoffs) or too low wages (due to, for instance, involuntary part-time or inappropriate minimum wages). Viewed from the demand side, it is especially seasonality and business cycles fluctuations that cause earnings insecurity. For the latter case, the TLM approach favours short-time work as a device for employment security cum partial earnings insurance.

Short-time work was extensively used in some EU countries during the 2008/09 recession (Eurofound 2010). Although often subsumed under *'passive'* labour market policy, the German case in particular demonstrates that short-time work (*Kurzarbeit*) is far from being 'passive' (Möller 2010). Instead, *Kurzarbeit* is consistent with the TLM approach if some of its risk-sharing elements were enhanced (Schmid 2015, pp. 84–86). In practice, however, short-time work has clear disadvantages compared to external flexibility covered by unemployment

insurance (UI). State subsidies may shift the costs to taxpayers or to marginal workers in favour of insiders; job security may maintain non-competitive industrial structures and lead to jobless growth or new job creation only in a non-standard form, especially temp-agency work. In implementing short-time work, Germany, for instance, failed in at least two respects from a TLM point of view: the incentives for training during short-time work are too low; and a corresponding flexible training infrastructure is still missing. All in all, however, the balance is positive even if it could be improved by complementary policies, especially lifelong learning.

Apart from providing earnings security in internal labour markets, risks related to job-to-job transitions ('external flexibility') have also to be taken into account. External employment mobility is often the only way to solve problems related to structural change due to new technologies and new product markets (see also next section). One important way to deal with the related earnings risks is to ensure transferability of social security entitlements across firms, industrial branches, regions or even – in the EU context – across countries. Excellent transport infrastructure – an example for embedded flexibility – is another way to allow workers to move to the places where labour demand is higher and where they feel most satisfied. Financial mobility incentives, including support for home equity insurance (Shiller 2003, p. 118f), might be an important element of 'active' securities where employment is hit by structural change and job creating investments at the former location are lacking.

Managing Transitions Between Employment and Unemployment

The biggest conventional risk during the life course is still the total loss of earnings through unemployment. This risk typically increases – beyond demand related causes – with unfavourable individual traits like low skills, migrant background or old age. The main question is to what extent this risk should be covered by collective or private insurance. TLM offers a clear answer to this question.

The origin of TLM's development was closely related to the erosion of the internal labour market (Doeringer and Piore 1971) which has been an implicit insurance contract: employers used to offer the male breadwinner a family wage, job security and lifelong earnings stability in exchange for the acceptance of wages below the productivity level at the peak of the work career. This implicit insurance contract targeted at male core workers was breaking down without a clear alternative in sight. Many countries with strong internal labour markets (France and Germany) bridged this institutional gap for a while with early retirement schemes and other 'golden handshake' practices. Unemployment insurance, in particular, was made instrumental as a 'social bridge' to retirement. When this led to an overburdening of social insurance finances, many European governments introduced stricter work conditionality ('activation') and included more 'employable' people into their UI systems, thereby levelling to some extent the benefits for all risk categories (Clasen and Clegg 2011, pp. 333–45). Many also transformed their pension insurances from benefit-based to contribution-based systems and from pay-as-you-go systems to funded systems. In this wake, two basic alternatives are under debate or are being practised: either private insurance elements like individual savings accounts and privately funded retirement schemes are being extended, or the new risks are included into universal social insurance. TLM theory argues for the second alternative, yet with some modifications allowing greater individual autonomy, which are discussed at length elsewhere (Schmid 2008, pp. 231–5 and 214–9; Schmid 2015).

As with all insurance, however, there is a trade-off. On the one hand, insurance has productive functions. People protected by social insurance engage in risky and profitable activities they would not otherwise undertake. Risky occupations might not be chosen without the protection of the welfare state, and it would be difficult to find entrepreneurs willing to undertake a risky investment if a debtor's prison were all that society provided should the venture fail. Eventually, employment-related insurance entitlements turn

out to be forceful incentives to find gainful employment, to moderate wages and to undertake training (Kolm and Tonin 2012). Without insurance employers might hesitate to restructure their production portfolio and workplaces if they were unsure of how to compensate their workers whose high firm-specific skills had reduced their employability for the external market. On the other hand, workers who know that a generous wage replacement is waiting if they quit their jobs (there may be a waiting period) might not bother to invest in firm-specific skills; they might even provoke dismissal or make a deal with their employer and fully exploit their entitlements. And by trusting in the social safety net, employers might also take recourse to dismissals instead of investing in the employability of their workforce. This is the moral hazard which preoccupies the majority of mainstream economists, forgetting the productive role social insurance can play, including properly designed unemployment insurance (Acemoglu and Shimer 2000).

How to balance productive and destructive risk-taking in a way that maximises equity and efficiency is an old conundrum of welfare state theory. If people choose more risks *ex ante*, they typically will be more unequal *ex post*. Risk-averse societies may exhibit relatively little inequality, but there is also little economic dynamism. By contrast, risk-taking societies may indeed exhibit high economic prosperity at the cost of high inequality, as the liberal US regime seems to show. Denmark, the 'flexicurity' model par excellence however, has received plaudits for reconciling high risk-taking and low inequality before *and* after taxes (Madsen 2006). It therefore does not seem that social insurance necessarily drives the 'big trade off between equality and efficiency' (Okun 1975); under certain circumstances it may well also drive a 'virtuous marriage between equality and efficiency' (Schmid 2008, pp. 314–322).

From the TLM point of view, it is a great mistake to speak of UI as a 'passive' measure. On the contrary, as long as the design of contributions and benefits respects the possible negative sides of any insurance, generous unemployment benefits (UB) are 'active' in the sense that they are an investment both in encouraging risk taking as well as in productive job searching, moreover, time-limited generous wage replacements are also a fair compensation for people who become unemployed through no fault of their own. Recent studies – even from the OECD (2010) – demonstrate that the unemployed with generous wage replacements in the first 6–9 months find more productive jobs (higher wages) than the unemployed not covered by unemployment insurance or covered only by means tested benefits. Even more important: these jobs are more sustainable, which means that decent wage replacements mitigate revolving door effects, i.e. leaving the benefits system and returning soon or entering another benefit system such as health or disability insurance (e.g. Wulfgramm and Fervers 2012).

'Active' labour market policy in conventional terms, i.e. placement or matching services, subsidised employment, training or even selected public works programmes to support transitions from unemployment back to employment, also play an important role in the TLM approach. Sustainable employment and enhanced employability are at the centre instead of workfare measures that emphasise the disciplinary function of 'activation' (de Koning 2007). Also in contrast to conventional evaluation research, implementation of 'active' labour market policy is at the fore (Schmid et al. 1996), for the simple reason that even highly sophisticated econometric evaluations usually provide only average impact measures that neglect potential positive effects through intelligent implementation even if – on the average – the overall effect is negative, therefore having very limited value for policy devices. Because most conventional evaluation studies limit the impact to a short period of time (one or at most 2 years after the measure), they neglect the possible long-term (career oriented) impacts which are central to TLM. Recent meta-evaluations, for instance, have shown that the long-term impact of education and training measures for the unemployed is much more favourable than the short-term impact, which may even be negative (Card et al. 2010).

Moreover, TLM theory insists on the need for local or regional negotiations, covenants or

employment pacts organising a wide set of transitions, in particular related to job-to-job transitions that prevent mass unemployment, a point emphasised early by Gazier (2002, pp. 222–7). A canonical example is the institution of *'work foundation'* in Austria, of which the *Voest-Alpine-Stahlstiftung* can be regarded as the prototype. These 'work foundations' serve as a kind of *transition agency* offering workers – who otherwise would have to be dismissed – an intermediate status of active 'job searchers'; the central aim is to maintain and enhance their employability and to find new jobs as quickly as possible. Several studies in the spirit of TLM addressed the issue of job-to-job transitions for preventing unemployment in cases of large-scale restructuring (van der Pas Borghouts- 2012; Bruggeman et al. 2012).

Managing Transitions Between Household Work and Employment

Managing the risk of restricted earnings capacity related to transitions from gainful employment to unpaid family work or – perhaps more relevant – combining both without undue loss of income, is another central aim of TLM theory. In practice, this risk is closely related to part-time work which has increased in much of Europe in recent decades. Although the dynamic has slowed down in recent years, part-time is still the most important form of non-standard employment, and recent studies came to the provocative result that increasing working time flexibility over the life course may not only be necessary for a decent work-life balance but also for high productivity in the knowledge and service economy (Schmid 2016; Schmid and Wagner 2017). Managing this transition dynamic in a way that marries both efficiency and equity considerations is a complex task because it affects not only employment or the labour market but also a range of other social policy issues like childcare, elderly care, health and old age insurance.

Inclusion of part-time work into unemployment insurance is quite common yet ensures only pro rata the reduced wage income due to part-time work. Income loss caused by transiting from full-time to part-time, due, for instance, to parental leave, has so far not been covered in most European countries. In Germany, for instance, the new parental leave allowance ('Elterngeld'), introduced in 2007, now insures the income loss due to full-time or part-time leave, as in the case of 'full-time' unemployment, by 67% of the former net wage income.

Involuntary part-time, however, is usually not covered but could help to make transitions pay: in many cases part-time serves as a stepping stone to full-time, and part-time unemployment insurance would provide an incentive for the unemployed to work part-time. It would also encourage employers to use a part-time job as a basis to test the employability of the unemployed. Moreover, Denmark and Sweden provide UI for involuntary part-time workers (according to MISSOC, Comparative Tables, July 2014), and the interim allowance (*Zwischengeld*) in Switzerland is a functional equivalent that insures the income gap between 'full-time' UB and the income of the new job (Schmid 2011, pp. 129–130).

A much-neglected opportunity would be the easy transition from full-time to (temporary) voluntary part-time and to provide part-time unemployment benefits under the condition that the other part of the 'working' time is used – apart from care work – for labour market education or training. Another option could be to reduce working time and to return later on to full-time. Both are, in principle, provided for in Germany but not much used owing to the prohibitive costs related to flexible work organisation and the fact that the right to return to full-time (at comparable conditions to those that existed before going part-time) cannot yet be properly enforced. Apart from parental part-time leave, the risk related to the reduced working time has to be shouldered completely by the individual if the labour law does not provide a helping hand, e.g. an obligation on employers to accept requests for temporary part-time unless he or she has good reason not to do so.

It is a well-established fact that equal tax treatment for married women has a strong positive effect on female labour force participation. In many EU countries married women, especially if they work part-time, are taxed more heavily than men or single women. A study of

17 OECD countries, for instance, shows that women will participate more when they are being taxed individually and equally compared to men (Jaumotte 2003). Germany still has joint taxation which heavily subsidises traditional partnerships (men as full-time wage earners, women – if any – only as marginal part-timers) and thus discourages women from increasing their involvement in paid employment and establishing their own social protection in old age. By contrast, Sweden is a good example of a system where the transfer from joint to separate taxation in combination with other family-friendly policies has led to higher labour force participation among women.

Finally, the importance of the state as employer not only of last resort but also as employer and promoter of public goods and services should not be neglected. High inclusive quality care or education is a collective action problem which the market does not solve, or solves only insufficiently (Atkinson 2015, pp. 140–47; Gottschall et al. 2015). The same holds true for providing adequate childcare in the spirit of making the market fit to workers. Here, equity and efficiency considerations open up a win-win situation: women's improved education can only be turned into productive capabilities if the tasks related to societal reproduction are solved through collective action. Last but not least, such a development would also facilitate the sharing of care responsibilities between men and women.

Managing Transitions Between Employment and Retirement or Disability

The fifth (expanding) cluster of risks is decreasing (or reduced) earning capacities over the life course due to health problems related to ageing, work related accidents or demanding work intensity. Only 11 of the 27 EU countries had reached the target of a 50% employment rate for senior people (age 55–64) by 2009. EU strategy nevertheless calls for a further increase of the retirement age, notably as a response to the demographic challenge (European Commission 2010). 'Active ageing', therefore, has become an accepted strategy (Hartlapp and Schmid 2008). Activity rates of older people differ markedly within the EU (or the OECD), and it is evident that varieties in pension schemes and in labour markets situations account for these differences (OECD 2016).

Little attention, however, has been given to the increased risk of reduced work capacities that often comes with ageing. For instance, evidence from the Nordic countries suggests that absence due to sickness is positively correlated with labour force participation (Schmid 2008, p. 159).

The TLM approach draws attention to the direct or indirect employment impact of social insurance systems not directly related to the labour market, such as, for instance, pension systems. Governments are now putting more emphasis on the 'activation functions' of such systems rather than – as at the end of the twentieth century – seeing, using or even exploiting them as escape routes for deficits in the systems of labour market insurance. It is not by accident that in many countries the extensive use of part-time work finds its parallel in a citizenship basic income in old age which is independent from the work history of a person, for instance in Switzerland, Denmark, Sweden, and particularly in the Netherlands (Visser 2003). In 2002, social partners in Finland – to mention just one example – reached an agreement on extensive pension reforms, thereby introducing explicit elements of employment insurance. People can retire flexibly at the age of 62–68. In this period, the accrual rate increases from its standard rate of 1.5% per year to 4.5 for those aged 63–68; the pension entitlements are no longer calculated on the income of the last 10 years but are instead based on lifetime earnings; and the scheme is automatically adjusted to the change in life expectation (Hartlapp and Schmid 2008).

Other sensible measures directed to the risk of reduced work (and related earnings) capacities are wage insurance or gradual retirement combining part-time work with part-time pensions. Of special importance to react properly to the need for flexible transitions into retirement are the establishment of the right to rehabilitation and, in particular, the right of employees and the obligation on employers to reasonable adjustment of workplaces according to reduced capacities of earnings (Deakin 2009).

The TLM approach also emphasises the value of negotiated flexibility and security beyond the establishment of legal rights and their reinforcement. Because employers' and employees' interests often do not converge, compromises have to be negotiated and implemented. Collective bargaining and agreements are often how such deals are attained. A good practice example for such coordinated flexibility is the German collective agreement established in the chemical industry in April 2008, setting up so-called demography funds. This overall framework agreement requires all employers to contribute an annual sum of €338 for each employee into a fund, which can be utilised, among other purposes, for training or retraining, for buying occupational disability insurance or for early retirement, while also facilitating the entry of young workers into employment. The recent collective agreement in this sector (27 March 2015) provides a stepwise increase of the amount to €750 in April 2017, which corresponds to an (otherwise) 0.9% increase in wages.

Concluding Policy Remarks

The central policy message of the TLM paradigm is that many of the new labour market risks go beyond unemployment for which UI was originally established. This development has led many countries to extend the spectrum of risks included into their social insurance – within or complementing their UI system. The TLM approach argues that it is high time to go a step further. There is a need for a systematic shift from simply insuring unemployment towards a system of employment insurance that covers risks beyond unemployment, in particular risks related to critical transitions over the life course: transitions between full-time and part-time work, transitions between one occupation and another, transitions between care work and gainful employment and transitions between full work-capacities and partial work-capacities. Many of these transitions can or could be organised within stable employment relationships thereby avoiding the exclusionary tendencies of non-standard employment. The theory of TLM aims to provide at least some strategic elements for such a paradigm shift which can be summarised in four points.

The first element is to establish a *general labour force membership status* through universal social rights and duties that include all kinds of work, paid *and* unpaid. Examples are the right to change working time over the life course, i.e. moving back and forth between full-time and part-time, and the right to some replacement of lost income due to reduced working time, especially in cases where the reduced income capacity is due to unpaid care work.

The second element is to establish a *career orientation* over the life course through *making transitions pay* and insuring life course risks beyond the risk of unemployment. The most promising example is public support of lifelong learning, especially (but not exclusively) for the low skilled; the benefit to the society would be enhanced mobility, in particular in the form of mobility chains that open up new ports of entry for outsiders. Another example would be wage insurance, providing some income protection when a change from a higher to a lower paid position arises.

The third element is to overcome inequalities and risk aversion through *capacity building*, for instance, through stepping stones, reasonable adjustment of workplaces, and active securities like drawing rights for investing in human capital. With this perspective, unemployment benefits have to be considered as 'active' and not as 'passive' security: as an investment in the search capacity of individuals and in the matching capacity of the labour market.

The fourth element is to *transform danger into trust* through negotiated flexibility and security, in particular through establishing learning communities in which not only social partners but also other regional key actors agree on specific employment objectives including not only job creation but also issues of lifelong learning and the humanisation of workplaces. Obviously, revitalising and renewing 'industrial relations' would be key in establishing an equitable and efficient system of employment insurance.

See Also

- European Labour Markets
- Social Insurance
- Social Insurance and Public Policy

Bibliography

Acemoglu, D., and R. Shimer. 2000. Productivity gains from unemployment insurance. *European Economic Review* 44: 1195–1224.

Anxo, D., C. Erhel, and J. Schippers, eds. 2007. *Labour market transitions and time adjustment over the life course*. Amsterdam: Dutch University Press.

Anxo, D., G. Boch, and J. Rubery. 2008. *The welfare state and life transitions: A European perspective*. Cheltenham/Northampton: Edward Elgar.

Atkinson, A.B.. 2015. *Inequality – What can be done?* Cambridge, MA/London: Harvard University Press.

Auer, P., and B. Gazier. 2006. *L'introuvable sécurité de l'emploi*. Paris: Flammarion.

Auer, P., J. Berg, and I. Coulibaly. 2005. Is a stable workforce good for the economy? Insights into the tenure-productivity-employment relationship. *International Labour Review* 144 (3): 319–343.

Borghouts-van der Pas, I. 2012. *Securing job-to-job transitions in the labour market – A comparative study of employment security systems in Europe*. Tilburg: W.L.P. (Wolf Legal Publishers).

Bruggeman, F., B. Gazier, and D. Paucard. 2012. Affronter les restructurations d'entreprise en Europe – Propositions pour une démarche unifiée. *Revue de l'IRES* 72: 29–64.

Brzinsky-Fay, C. 2011. *School-to-work transitions in international comparison*, 1663. Tampere: Acta Universitatis Tamperensis.

Burroni, L., A. Lobascio, and M. Pedaci. 2010. *The regional governance of economic uncertainty*. GUSTO working paper WP6. www.gusto-project.eu

Calandrino, M., and S. Gagliarducci. 2004. Labour market transitions and advancement: Temporary employment and low-pay in Europe. In *European Commission, Employment in Europe 2004*, 150–186. Brussels: DG for Employment and Social Affairs.

Card, D., J. Kluve, and A. Weber. 2010. Active labour market policy evaluations: A meta-analysis. *The Economic Journal* 120 (548): F452–F477.

Clasen, J., and D. Clegg, eds. 2011. *Regulating the risk of unemployment – National adaptations to post-industrial labour markets in Europe*. Oxford: Oxford University Press.

De Koning, J., ed. 2007. *Evaluating active labour market policy – Measures, public private partnerships and benchmarking*. Cheltenham: Edward Elgar.

Deakin, S. 2009. Capacitas: Contract law, capabilities and the legal foundations. In *Capacitas – Contract law and the institutional preconditions of a market economy*, ed. S. Deakin and A. Supiot, 1–29. Oxford/Portland: Hart Publishing.

Doeringer, P.B., and M.J. Piore. 1971. *Internal labor market and manpower analysis*. Lexington: D. C. Health and Company.

Dworkin, R. 2000. *Sovereign virtue. The theory and practice of equality*. Cambridge, MA/London: Harvard University Press.

Ebner, C. 2012. *Erfolgreich in den Arbeitsmarkt? Die duale Berufsausbildung im internationalen Vergleich*. Frankfurt a.M./New York: Campus.

Eichhorst, W., N. Rodríguez-Planas, R. Schmidl, and K.F. Zimmermann. 2012. *A roadmap to vocational education and training systems around the world*. IZA discussion paper no. 7110. Bonn.

Eurofound. 2010. *Extending flexicurity – The potential of short-time working schemes*. Dublin: European Foundation for the Improvement of Living and Working Conditions.

Eurofound. 2012. *NEETs: Young people not in employment, education or training: Characteristics, costs and policy responses in Europe*. Luxembourg: Publications Office of the European Union.

European Commission. 2010. An agenda for new skills and jobs: A European contribution towards full employment. Communication from the Commission, COM(2010) 682 final, Strasbourg, 23 Nov 2010.

European Commission. 2012. New skills and jobs in Europe: Pathways towards full employment. Office for Official Publications of the European Communities (Report written by Günther Schmid), Luxembourg. http://ec.europa.eu/research/social-sciences/pdf/new-skils-and-jobs-in-europe_en.pdf

Eurostat. 2016. Labour market flow statistics in the EU, Eurostat. http://ec.europa.eu/eurostat/statistics-explained/index.php/Labour_market_flow_statistics_in_the_EU. Downloaded 5 Nov 2011.

Gazier, B. 2002. Transitional labour markets: From positive analysis to policy proposals. In *The dynamics of full employment – Social integration through transitional labour markets*, ed. G. Schmid and B. Gazier, 196–233. Cheltenham/Northampton: Edward Elgar.

Gazier, B. 2003. *'Tous Sublimes' – Vers un nouveau plein-emploi*. Paris: Flammarion.

Gazier, B., and J. Gautié. 2009. *The "transitional labour markets" approach: Theory, history and future research agenda*. Paris: Documents de Travail du Centre d'Économie de la Sorbonne. https://halshs.archives-ouvertes.fr/halshs-00363404.

Gottschall, K., B. Kittel, K. Briken, J.-O. Heuer, S. Hils, and M. Tepe. 2015. *Public sector employment regimes – Transformation of the state as an employer*. Houndmills/New York: Palgrave Macmillan.

Guergoat-Larivière, M., and C. Erhel. 2010. *Labour market status, transitions and gender: A European perspective*. Paris: Documents de Travail du Centre d'Économie de la Sorbonne. https://halshs.archives-ouvertes.fr/halshs-00484577.

Hartlapp, M., and G. Schmid. 2008. Labour market policy for, active ageing in Europe: Expanding the options for retirement transitions. *Journal of Social Policy* 37(3): 409–431.

Heckman, J.J. 2008. Schools, skills, and synapses. *Economic Inquiry* 46 (3): 289–324.

Hirschman, A.O. 1970. *Exit, voice and loyalty – Responses to decline in firms, organizations and states*. Cambridge, MA: Harvard University Press.

Jaumotte, F. 2003. *Female labour force participation: Past trends and main determinants in OECD countries*. OECD Economics Department working papers, no. 376. Paris.

Jørgensen, H., and P.K. Madsen, eds. 2007. *Flexicurity and beyond – Finding a new agenda for the European social model*. Copenhagen: DJØF Publishing.

Kahneman, D. 2011. *Thinking fast and slow*. London: Allan Lane (Penguin Books).

Kolm, A.-S., and M. Tonin. 2012. *In-work-benefits and the Nordic model*. IZA discussion paper no. 7084. Bonn.

Korver, T., and P.R.A. Oeij. 2008. Employability through covenants: Taking external effects seriously. In *The European social model and transitional labour markets – Law and policy*, ed. R. Rogowski, 143–169. Farnham/Burlington: Ashgate.

Korver, T., and G. Schmid. 2012. Enhancing transition capacities and sustainable transitions. In *Renewing democratic deliberation in Europe: The challenge of social and civil dialogue*, eds. J. de Munck, C. Didry, I. Ferreras and A. Joberts, 23–55. Brussels: Peter Lang.

Lassen, M., J.H. Sørensen, A. Lindkvist Jørgensen, and R.J. Møberg. 2006. *Skill needs and the institutional framework conditions for enterprise-sponsored CVT – The case of Denmark*. WZB-discussion paper SP I 2006-121. Berlin.

Lassnigg, L., H. Burzlaff, M.A.D. Rodriguez, and M. Lassen, eds. 2007. *Lifelong learning – Building bridges through transitional labour markets*. Apeldoorn/Antwerpen: Het Spinhuis.

Leschke, J. 2015. Non-standard employment of women in service sector occupations: A comparison of European countries. In *Non-standard employment in post-industrial labour markets: An occupational perspective*, ed. W. Eichhorst and P. Marx, 324–352. Cheltenham/Northampotn: Edward Elgar.

Madsen, P.K. 2006. How can it possibly fly? The paradox of a dynamic labour market. In *National identity and the varieties of capitalism – The Danish experience*, ed. J.L. Campbell, J.A. Hall, and O.K. Pedersen, 321–355. Montreal: McGill-Queen's University Press.

Möller, J. 2010. The German labor market response in the world recession – De-mystifying a miracle. *Journal for Labour Market Research* 42 (4): 325–336.

Muffels, R.J.A., ed. 2008. *Flexibility and employment security in Europe – Labour markets in transition*. Cheltenham/Northampton: Edward Elgar.

O'Reilly, J., ed. 2003. *Regulating working-time transitions in Europe*. Cheltenham/Northampton: Edward Elgar.

OECD. 2010. Institutional and policy determinants of labor market flows. In *OECD-Employment Outlook 2010*, 167–210. Paris.

OECD. 2016. *Pensions at a glance 2015*. Paris: OECD Publications.

Okun, A.M. 1975. *Equality and efficiency – The big trade-off*. Washington, DC: The Brookings Institution.

Rawls, J. 2001. In *Justice as fairness – A restatement*, ed. Erin Kelly. Cambridge, MA/London: The Belknap Press of Harvard University Press.

Rogowski, R. 2008. The European social model and the law and policy of transitional labour markets in the European Union. In *The European social model and transitional labour markets – Law and policy*, ed. R. Rogowski, 9–27. Farnham/Burlington: Ashgate.

Sabel, C.F. 1994. Learning by monitoring: the institutions of economic development. In *Rethinking the development experience: Essays provoked by the work of Albert O. Hirschman*, ed. L. Rodwin and D.A. Schön, 231–274. Washington, DC/Cambridge, MA: The Brookings Institutions and The Lincoln Institute of Land Reform.

Schmid, G. 1995. Is full employment still possible? Transitional labour markets as a new strategy of labour market policy. *Economic and Industrial Democracy* 16 (3): 429–456.

Schmid, G. 2002. *Wege in eine neue Vollbeschäftigung – Übergangsarbeitsmärkte und aktivierende Arbeitsmarktpolitik*. Frankfurt a.M./New York: Campus Verlag.

Schmid, G. 2008. *Full employment in Europe – Managing labour market transitions and risks*. Cheltenham/Northampton: Edward Elgar.

Schmid, G. 2011. *Übergänge am Arbeitsmarkt – Arbeit, nicht nur Arbeitslosigkeit versichern*. Berlin: edition sigma.

Schmid, G. 2014. Transitional labour markets and employment services. In *Building bridges – Shaping the future of public employment services towards 2020*, ed. F. Leroy and L. Struyven, 71–100. Brugge: die Keure Professional Publishing.

Schmid, G. 2015. Sharing risks of labour market transitions: Towards a system of employment insurance. *British Journal of Industrial Relations* 53(1): 70–93. http://www.guentherschmid.eu/pdf/Sharing_Risks_BJIR-2015.pdf

Schmid, G. 2016. *Flexible and secure labour market transitions: Towards institutional capacity building in the digital economy*. IZA policy paper no. 116. Bonn. http://ftp.iza.org/pp116.pdf

Schmid, G., and B. Gazier, eds. 2002. *The dynamics of full employment. Social integration through transitional labour markets*. Cheltenham/Northampton: Edward Elgar.

Schmid, G., and J. Wagner. 2017. *Managing social risks of non-standard employment in Europe*. ILO-working paper, Conditions of Work and Employment Series no. 91.

Schmid, G., J. O'Reilly, and K. Schömann, eds. 1996. *International handbook of labour market policy and evaluation*. Cheltenham: Edward Elgar.
Schömann, K., and P.J. O'Connell, eds. 2002. *Education, training and employment dynamics – Transitional labour markets in the European Union*. Cheltenham: Edward Elgar.
Sen, A. 2009. *The idea of justice*. London: Allan Lane and Penguin Books.
Shiller, Robert J. 2003. *The new financial order – Risk in the 21st century*. Princeton: Princeton University Press.
Supiot, A. 2001/2016. *Beyond employment: Changes in work and the future of labour law in Europe*. Oxford: Oxford University Press; second edition with a new preface in French by Alain Supiot.
Van den Berg, A., and E. de Gier. 2008. Research in transitional labour markets: Implications for the European employment strategy. In *The European social model and transitional labour markets – Law and policy*, ed. R. Rogowski, 63–105. Farnham/Burlington: Ashgate.
Visser, J. 2003. Negotiated flexibility, working-time and transitions in the Netherlands. In *Regulating working-time transitions in Europe*, ed. J. O'Reilly, 123–169. Cheltenham/Northampton: Edward Elgar.
Wulfgramm, M., and L. Fervers. 2012. *Unemployment and subsequent employment stability: Does labour market policy matter?* IZA DP no. 7193. Bonn.
Zimmermann, B. 2011. Making employees' pathways more secure: A critical examination of the company's responsibility. In *Transforming European employment policy – Labour market transitions and the promotion of capabilities*, ed. R. Rogowski, R. Salais, and N. Whiteside, 117–137. Cheltenham/Northampton: Edward Elgar.

Transitivity

Wayne Shafer

Keywords

Convexity; Integrability; Law of demand; Principle of persistent nonpreference; Strong axiom of revealed preference; Transitivity; Weak axiom of revealed preference

JEL Classifications

C0

Transitivity is formally just a property that a binary relation might possess, and thus one could discuss the concept in any context in economics in which an ordering relation is used. Here, however, the discussion of transitivity will be limited to its role in describing an individual agent's choice behaviour. In this context transitivity means roughly that if an agent chooses A over B, and B over C, that agent ought to choose A over C, or at least be indifferent. On the surface this seems reasonable, even 'rational', but this ignores how complicated an agent's decision making process can be. For an excellent discussion of this issue see May (1954). Given a model of agent behaviour, transitivity can be imposed as a direct assumption, or can be an implication of the model for choice behaviour. The standard model of agent behaviour in economics is that the agent orders prospects by means of a utility function, which in effect assumes transitivity. With appropriate continuity and convexity restrictions on utility functions, the model allows one to demonstrate that: (1) Individual demand functions are well defined, continuous, and satisfy the comparative static restriction, the strong axiom of revealed preference (SARP). (In the smooth case, this corresponds to the negative semi-definiteness and symmetry of the Slutsky matrix.) (2) Given a finite collection of such agents with initial endowments of goods, a competitive equilibrium exists. What will be discussed in the remaining part of this article is to what extent one can obtain results analogous to (1) and (2) above while using a model of agent behaviour which does not assume or imply transitive behaviour. To keep the discussion as simple as possible, we will only consider the situation in which the agent's set of feasible commodity vectors is the non-negative orthant of n-dimensional Euclidean space, and the agent's problem is to choose a commodity vector x when faced with positive prices and income. A vector p in the positive orthant of Euclidean n-space will denote the vector of price-income ratios, or a 'price' system.

Two models of agent behaviour which have a long history in economics will now be described. The first, which will be called the 'local' theory, takes as its primitive the assumption that if an

agent is currently consuming at a vector x, he is able to determine if an infinitesimal change in x, say x to $x + dx$, is a change for better or worse. This idea is represented by a function $x \rightarrow g(x)$, mapping each vector x into a n-vector $g(x)$ such that a small movement from x in the direction of y is an improvement if $g(x)(y - x) > 0$, and not an improvement otherwise. Given a price system p, an affordable x is an equilibrium point for the agent if $g(x)(y - x) \leq 0$, for all affordable y. That is, no small movement from x in the direction of an affordable y is an improvement. A basic question is whether for every p, an equilibrium x exists. This approach goes back at least to Pareto, and most economists, including Pareto, concerned themselves with the 'integrability' problem: when is there a quasi-concave utility function such that $g(x)$ is a positive scalar multiple of the vector of marginal utilities, for each x? Note that if an agent has a differentiable quasi-concave utility function u, and one defines g by $g(x) = \lambda(x)Du(x)$ for any $\lambda(x) > 0$ then $g(x)(y - x) > 0$ is equivalent to $Du(x)\ (y - x) > 0$, and this implies $u(x + t(y - x)) > u(x)$ for t positive and sufficiently small. Thus if g is 'integrable', the agent acts as if he maximizes a utility function, and thus results (1) or (2) above will be satisfied. Some economists, however, believed that the local theory could be used to describe agent behaviour without assuming the integrability conditions. Most notable is the work of Allen (1932), Georgescu-Roegen (1936, 1954) and Katzner (1971). Without the integrability conditions and the implied utility function (and thus implied transitivity) the existence of an equilibrium x given any p is nontrivial. This problem was solved by Georgescu-Roegen (1954), who showed that if g is continuous and g satisfied the 'principle of persistent nonpreference' (PPN), that is, $g(x)(y - x) < 0$ implies $g(x)(y - x) > 0$, then an equilibrium point will exist in any budget set. It should be noted that the integrability problem mentioned above requires PPN (for quasi-concave utility), as well as the Frobenius conditions for mathematical integrability. It is easy to show that Georgescu-Roegen's assumptions imply that the resulting demand correspondence will be upper-hemicontinuous.

The second basic approach to modelling agent behaviour will be called the 'global' theory. In this approach, the primitive of the theory is a binary relation R on the commodity space with xRy having the interpretation 'x is at least as good as y'. Define the strict preference relation P by xPy is equivalent to not yRx. (P could also be taken as the primitive.) Given a price system p, an affordable x in an equilibrium point if yPx implies $py > 1$, that is, any vector y preferred to x is not affordable. A basic question is whether such an equilibrium point will exist. This approach dates back to Frisch (1926), and the usual approach was to specify conditions on R which imply R has a representation by a continuous utility function, that is, $u(x) \geq u(y)$ equivalent to xRy. This problem was solved by Debreu (1954), who showed that R must be reflexive, complete, transitive and continuous. With the addition of appropriate convexity conditions, this approach yields the results (1) and (2) above. However, in a remarkable paper, Sonnenschein (1971) showed that one could remove transitivity from the list of standard assumptions and still have a well defined demand correspondence which is upper-hemicontinuous. Specifically, he demonstrated that if R is continuous, reflexive, and $P(x) = \{y : yPx\}$ is convex for all x, then an equilibrium point will exist in any budget set, and the resulting demand correspondence is upper-hemicontinuous. (He also assumed R is complete, but that assumption was used only to show that an equilibrium point is comparable to every affordable y.) Note that if g represents local theory, and g is continuous, then the R defined by xRy is equivalent to $g(x)(y - x) \leq 0$, satisfies Sonnenschein's conditions, and an equilibrium x for R is an equilibrium point for g, in any budget. Thus Sonnenschein demonstrated that Georgescu-Roegen's condition PPN is not necessary for the existence of equilibrium points in a budget set.

In order to resolve question (1) above, a theory must predict a unique equilibrium point in each budget set, in order to get a well defined demand function. In the local theory, if one assumes $g(x) \neq 0$ for all x and strengthens PPN to SPPN: $g(x)(y - x) \leq 0$, $x \neq y$ implies $g(x)(y - x) > 0$,

then the equilibrium point x will be unique in any budget, and the resulting demand function will be continuous and satisfy the weak axiom of revealed preference (WARP). Thus the local theory, without assuming mathematical integrability (implied transitivity), yields a theory of individual demand functions satisfying WARP. On the other hand, given a continuous demand function h satisfying WARP, if h has a continuous inverse, then $g = h^{-1}$ yields a local theory with g satisfying SPPN and generating h. Now consider the global theory. If R is represented by a continuous, strictly quasiconcave nonsatiated utility function, then R will be reflexive, complete, transitive, strongly convex and nonsatiated. If one simply removes the assumption of transitivity from this list, then Sonnenschein's result implies that an equilibrium point will exist, and the remaining assumptions imply that this point will be unique, and that the resulting demand function will be continuous and satisfy WARP (see Shafer 1974). Furthermore, Kim and Richter (1986) showed that, with a slight variation in the assumptions on R, any continuous demand function h satisfying a modified version of WARP can be generated by such an R. Thus, from the point of view of having single valued demand functions, the absence of transitivity, either assumed as in the global theory or implied as in the local theory, is essentially equivalent to Samuelson's theory of observed demand satisfying WARP. Since WARP includes the 'law of demand', that is, normal goods have downward sloping demand, in my view little is lost by not assuming transitivity.

Now question (2) above, the problem of existence of competitive equilibrium will be discussed. Again, Sonnenschein observed that if one took the standard assumptions on R normally used in proofs of existence of a competitive equilibrium, and removed the transitivity assumption, then demand correspondences would be well defined and convex valued, and the standard proof techniques would still work, so equilibrium would exist. Thus transitivity is irrelevant to demonstrating the internal consistency of the competitive model. Note, however, that the assumptions needed by Sonnenschein to demonstrate that individual demand correspondences are well defined and upper-hemicontinuous, namely continuity and convex preferred sets, are too weak to ensure convex valued demand correspondences. Nevertheless, Mas-Colell (1974) demonstrated that with only these assumptions on preferences, competitive equilibria will exist. Thus the only properties of individual preferences which are important to the existence of competitive equilibrium are continuity and convexity.

See Also

- ▶ Integrability of Demand
- ▶ Orderings

Bibliography

Allen, R.G.D. 1932. The foundations of a mathematical theory of exchange. *Economica* 12: 197–226.

Debreu, G. 1954. Representation of a preference ordering by a numerical function. In *Decision processes*, ed. R.M. Thrall, C.H. Combs, and R.L. Davis. New York: Wiley.

Frisch, R. 1926. Sur un problème d'économie pure. *Norsk matematisk forenings skrifter* 16: 1–40.

Georgescu-Roegen, N. 1936. The pure theory of consumer's behavior. *Quarterly Journal of Economics* 50: 545–593.

Georgescu-Roegen, N. 1954. Choice and revealed preference. *Southern Economic Journal* 21 (October): 119–130.

Katzner, D. 1971. Demand and exchange analysis in the absence of integrability conditions. In *Preferences, utility, and demand*, ed. J. Chipman et al. New York: Harcourt, Brace, Jovanovich.

Kim, T., and M. Richter. 1986. Nontransitive-nontotal consumer theory. *Journal of Economic Theory* 38: 324–363.

Mas-Colell, A. 1974. An equilibrium existence theorem without complete or transitive preferences. *Journal of Mathematical Economics* 1: 237–246.

May, K. 1954. Intransitivity, utility, and aggregation in preference patterns. *Econometrica* 22: 1–13.

Shafer, W. 1974. The nontransitive consumer. *Econometrica* 42: 913–919.

Sonnenschein, H. 1971. Demand theory without transitive preferences, with applications to the theory of competitive equilibrium. In *Preferences, utility, and demand*, ed. J. Chipman et al. New York: Harcourt, Brace, Jovanovich.

Transport

A. A. Walters

The obvious definition of 'transportation' is the movement of goods or people over space. But conventionally we do not include short trips inside the household or office or warehouse or factory as part of transport; such activities of movement are reckoned to be part of the household chore or industrial process. Only movements outside the home or factory are normally reckoned to require the services of the transport sector, as normally defined. In many definitions of 'transport', particularly in Western industrialized countries, non-motorized forms of carriage are excluded. Walking trips or manually hauled freight, bicycle journeys and even animal-powered journeys are usually excluded – except in those cases where walking may play a ubiquitous role, or in cities such as Amsterdam and Cambridge, where bicycles are a common form of transit. In third world countries, however, such manual or animal-powered operations still play a significant, perhaps a major role in the sector. Even in the mid-1980s it is very likely that, considering only passenger trips of over one kilometre in length, the vast majority in the third world are walked (World Bank 1985).

In measuring the size of the transport sector, these issues of definition must constantly be borne in mind. If we confine the transport sector to mechanical carriage, then it is likely that in most Western industrialized countries, transport accounts for about 10–12% of the gross domestic product (GDP). As well as excluding walking and cycling, this percentage does not take into account the value of time that people and goods spend in transit. If some allowance is made for this time (at values discussed below), then the percentage would rise to over 15%, and perhaps as much as 20%, of GDP. In primitive societies in the third world, such as the Sahel countries of Africa, the percentage is likely to be much higher, whereas in the constrained city states of Singapore and Hong Kong it is rather lower.

Throughout the world the dominant form of transport is by road. In the Western industrialized countries, some 75–85% of transport sector resources are attributed to truck, bus or car. The remaining 15–25% are distributed over rail, water and air, depending on geographical, historical and political conditions. In the third world this dominance of road transport, with one or two important exceptions such as India, tends to be even more pronounced. Transport by truck, bus and car is growing at rates higher than the growth of GNP in much of the third world.

Institutions and Organizations

The institutional structure of transport tends to be broadly similar in all countries. The road transport industry is typically organised in small, even very small, owner-driver units operating in a competitive, if often regulated, environment. Railways, *per contra*, are generally large monopolies almost always owned and, in principle at least, regulated by the state. State-owned airline monopolies are still the normal form of organization, the outstanding exceptions being the United States. But there is some evidence of a trend towards competition and private ownership in air transport. Ocean transport, especially the bulk-cargo business, is mainly in the hands of competitive private owners. Many countries have state-owned shipping lines, but although they often have reserved cargoes, they generally try to compete for free cargo.

The institutional and organizational structure of transport has affected the study of transport economics – and perhaps even the results of such studies have influenced government policies and so affected institutional forms. Transport economics has hardly been one of the central concerns of the discipline of economics. Yet over the years it has provided a fruitful field for the development of ideas which turn out to have a wide field of application.

Monopoly Power and Regulations

In the nineteenth century, apart from the interest in transport costs as some sort of natural barrier to

trade, the interests of economists centred mainly on the problems of controlling what were widely thought to be natural monopolies. Regulation of tariffs and fares pre-dated the railways, but the rapid expansion of steam locomotion raised the issue of control and regulation to the centre of political dispute.

From the middle of the nineteenth century, laws were passed prohibiting 'undue' (UK) or 'unfair' (USA) preference in the fixing of rail freight rates and fares. Behind this legislation lay the widespread fear that the railways would use their monopolistic bargaining powers to discriminate against the small shipper and *a fortiori* against the individual passenger. From the last quarter of the nineteenth century the hazy ideas on which this legislation was based were redefined by Marshall, Taussig, Edgeworth, Pigou and Ramsey. The theory of a discriminating monopoly provided a rationalization for 'charging what the traffic will bear', as well as providing the basis of modern theories of optimum taxation and utility pricing.

Besides requiring what were thought to be fair prices, the law had something to say about the provision of services. The general idea was to prevent the railways using withdrawal, or the threat of withdrawal, as a way of avoiding their common carrier obligation to provide adequate service. The implication was that the railways should maintain services even if that implied that loss-making services had to be cross-subsidized from profitable ones. Throughout the twentieth century the closure of a rail line or even the withdrawal of a service has entailed major political or legal action.

The railways were never the absolute monopolists of popular fable. Their power was much restricted by other rail companies, water transport and highway competition. Initially, unregulated trucks and buses grew rapidly after World War I and were said to be 'creaming' off the railways' profitable traffic of commodities with a high value relative to bulk. Thus the railways argued that they were left with the low-value, bulky traffic such as coal and minerals, which could not bear a high freight rate. The railways and established large truckers complained that the trucking industry, and particularly new entrants, were indulging in 'wasteful competition' with cut-throat tariffs. These complaints, in addition to excessive, repeated bankruptcies in the 1930s, led in Britain to the substantial control of entry and fares, and in the USA and Germany to trucking and bus tariffs.

In the 1930s, 1940s and much of the 1950s, transport economists were concerned largely with the examination of the consequences, and a judgement of the efficacy, of the restrictive rail legislation of the previous century and new regulatory mechanism imposed on trucks and buses. Gilbert Walker (UK), James Nelson (USA) and Walter Eucken (Germany) showed that the conventional wisdom was quite discredited by the evidence (Walker 1948). The common understanding that the railways made profits from their high-value traffic and lost money on their bulk traffic was quite wrong – the opposite of the truth. (In Gilbert Walker's aphorism, the railways' traffic cream was at the bottom, not the top, of the bottle.) Second, it was shown that the preregulation bankruptcy rates of truckers were unusually low, not high, and that regulation was primarily an obstacle to efficiency and innovation. Third, the scholars demonstrated that regulation entailed absurd wastes of resources in empty back-hauls, idle capacity and, in the United States, circuitous routings (*see* REGULATION AND DEREGULATION). They showed the unmistakable signs of what came later to be called the captive regulatory agency. (It was perhaps the sign of the times that they did not extend their analysis to the wastes that restrictive trades union behaviour so encouraged.) The rapid accumulation of evidence of the waste, inefficiency and inequity of regulation was the main factor behind the deregulation movement which gathered momentum in the 1950s and 1960s, culminating in the extensive deregulations of the 1970s and 1980s (Keeler 1983).

Although (in 1985) it is rather too soon to give a definitive view of the effects of deregulation, the interim result appear to be entirely consistent with the implications suggested by transport economists such as Walker and Nelson. No evidence of 'wasteful competition', whatever that omnibus term may mean, can be readily detected in the deregulated trucking industry (e.g. in the UK or

Australia). And freedom has meant a large increase in efficiency with lower fares, as in the United States airline industry. Significantly, there has been little evidence of any movement to re-regulate those industries which have been freed (Friedlander 1981).

The precise process of deregulation has varied considerably, depending on political and social conditions and whether or not the industry has substantial elements of public ownership. Most of the railroads of the world had been taken into public ownership by the 1950s, the most notable exception being those in the United States. Regulations requiring the privately owned rail companies to maintain unprofitable services, limitations on their pricing policies, and notorious featherbedding and restrictive practices had all served to render them financially nonviable and dependent on subsidies from the state. The flight of private capital from such an unpromising industry was countered by nationalization. With persistent regulation, however, changing the ownership did not have any substantial beneficial effect on what had come to be called the railway problem, except to make the railways more subject to political pressure and union militancy. The financial deficits waxed rather than waned – to become a major worry to finance ministers all over the world. The fear that such deficits would grow even faster if road transport were free to compete with rail has been one of the main constraints on deregulation of trucking and buses in continental Europe and Japan, thus providing an example of the chain reaction that is typical of regulation. It is unlikely that there will be any substantial privatization and deregulation of railways in the foreseeable future. On the other hand, because of the example of the United States, there does seem both scope and hope for substantial deregulation and privatization of the airline industry in Europe, together with some further freeing of trucks and buses.

Supply and Cost

From the 1950s onwards the content of transport economics changed considerably. Although work continued on the issues of law and regulation, the new transport economists were more interested in the analysis and, above all, the measurement of economic phenomena in transport. Analysis primarily took the form of seeking to refine and develop basic spatial models of economic activity and the specification of the salient characteristics of transport supply and demand.

From the *supply* side, one of the main aims has been to determine the structure of the production function and its dual, the cost function. In particular, the implementation of regulations often, implicitly at least, required evidence on costs. But the theoretical as distinct from the political rationalization for subsidies to rail, ports, airports and so on, required demonstration of substantial economies of scale. The earliest studies of rail costs focused on whether total costs varied less than proportionately to the variation of traffic. Lumpiness in the production of transport services and discontinuous jumps in the cost function had been recognized in the nineteenth century (Lardner 1850). The track is fixed and so its costs vary little with respect to traffic, until an additional track is needed. By plotting railway costs against ton-miles, the first studies suggested that there were considerable economies of scale in the average railway operation.

Perhaps the most important development in cost function analysis from the late 1950s onwards was the integration of the engineering and economic approaches (Wohl and Hendrickson 1984). Both models and data were taken over from engineering and interpreted in economic terms, often thereby improving the engineers' interpretation and forecasts. The engineering–economics applications were particularly fruitful in pipelines, airlines and shipping, but above all in the analysis of road traffic. For example, economic interpretation of the traffic engineers' gravity model provided new insights, both economic and engineering. Old data could be usefully reworded and reinterpreted. Perhaps the most important particular benefit was in the analysis of the social costs of congestion and the development of better systems of user costs for highways and urban streets. The various statistical models of the theory of queues and basic principles from the theory of fluid dynamics have

proved useful bases for modelling traffic flow, with results that have been interpreted in terms of social costs of congestion.

The specification of the transportation process in terms of its engineering elements, the attribution of unit costs to inputs and the development of a *synthetic cost function* have been particularly useful for modelling the multiproduct nature of transport services. The carriage of a consignment or passenger from point A to point B in a particular time interval is a service distinct from, and with finite substitutability for, another service from C to D in another time interval. Yet clearly *some* aggregation is required over space and time. Transportation firms themselves tend to think and to price their services in terms of aggregates, but in the econometric analysis such aggregation still tends to be rather *ad hoc*.

The basic theoretical requirement of aggregation – additive separability – does not readily apply to most transportation processes. Moreover, the variability of marginal costs due to the variations in load factor, or empty capacity, makes it particularly hazardous to aggregate outputs. (The marginal costs of freight when capacity is full may be more than three times the marginal costs where there is empty space to be filled on a back-haul.)

The levels of costs that have emerged from the econometric–engineering studies have been most useful for comparative purposes. For example, one important application is whether a rail, subway or bus system should be developed to deal with the increased demand for urban passenger transport. As shown in World Bank (1985), such comparative studies have shown that subway rail systems costs two or three times that of alternative bus systems.

The most important application of the cost model is the issue of the existence and extent of economies of scale. As the nineteenth-century studies suggested, railways experience substantial economies of scale; more recent studies have shown their cost elasticities ranging from 0.35 to 0.80. By using their fixed assets (particularly the track) more intensively, unit costs can be much reduced. In road transport, *per contra*, scale economies are quickly exhausted, and the trucking and bus firms soon reach their minimum-cost outputs. (Such results are consistent with the ubiquity of the small firm for both bus and trucking operations in an unregulated environment.) Similarly for airlines, the evidence suggests that, for the prevailing size of airline in the United States, there are no signs of scale economies.

Refinements of the production function, and in particular the cost function, have included the elaboration of the specification of qualitative characteristics of the output of services (so-called hedonic measurements) and the influence of the stochastic nature of demand on costs. It was always clear that quality of service played a dominant role in the valuation of transport services. Speed of delivery (freight) and time of trip (passenger) as well as comfort and safety were the necessary *ceteris paribus* of competitive studies of cost functions. But shipper and passenger valuations of such quality-of-service variables must be incorporated in cost models if the results are to have useful validity.

The value which is to be attributed to time is also a parameter of some, often critical, importance. Although time is often important even in freight transport, it is clearly of crucial importance in passenger transit. Many comparisons of cost also involve differences in the time of trip, so that some evaluation of time savings is necessary in order to draw implications. The market value of time can be observed as people trade off time against money in decisions such as the use of a toll road rather than a free highway, or in taking an air trip rather than a surface trip. From studies of these situations, the dispersions of the implicit values of time have been considerable, yet some signs of regularity have been observed (Winston 1985). Thus: (a) if the person is travelling on his firm's time, the savings of time can be valued at the wage costs per hour of the person involved; (b) if he is travelling during his leisure time and in a vehicle, then he values time saved at roughly one-third to one-quarter of his earnings during his working time; and (c) if he is waiting at the bus stop or rail platform, then he values such waiting time about as much as his wage rater per hour (World Bank 1985). These are the trade-offs observed in the market for time of trip, and in a

competitive system these trade-offs will be reflected in the decisions of competitive firms in their production functions.

Safety is yet another characteristic of transport services, again particularly passenger services, which is implicit but important. Indeed, traffic accidents are a major source of mortality in developing countries (World Bank 1985). The valuation of injury or loss of life is fraught with daunting philosophical difficulties. But fortunately we do not need to ask, let alone answer, the questions concerned with the 'value of a life'. In practice the person travelling makes an implicit decision about the change in the probability of losing his life (or being injured, etc.) and the cost involved. No one 'buys' a life, either his own or that of any other person, on the market (except for certain villains); but people do implicitly value *changes in the probability of being killed or injured*. This trade-off between safety and money and time is practised when a man jaywalks or when he skips a servicing of the brakes and steering of his motor car. Little research, however, has been pursued to discover these implicit valuations. Most of the valuations of life and limit which are employed in practical applications, such as highway construction and roaduse regulations, have been calculated from the money costs of damage, loss of useful output while incapacitated, the costs of medical services, and some notional sum for pain and discomfort.

Finally, there is the problem of allowing for other externalities or 'the social and environmental effects' of transport activities. Air pollution, noxious noise and the disfigurement of the landscape with perhaps the loss of part of the 'national heritage' are the most important items considered in such calculations. The best measure of the cost is the loss of rent of facilities, such as houses, shops, land, and so on, which are affected. This loss of rent needs to be inserted in a dynamic model of the adjustment process; then one can find the capitalized net present value of the disamenity created by the new transport facility. The most sophisticated applications of this environmental model have been for aircraft noise shadows associated with new runways at airports (Walters 1978). The valuation of loss of amenity is generally carried out by some version of the Clawson method.

The Demand for Transport

The reaction of people and firms to a reduction or increase of fares, transport costs or freight rates has always been a central issue of economic development. A general knowledge of what were later to be called demand elasticities has been incorporated in tariffs and fares since classical times. The principle of charging 'what the traffic will bear' and fixing freight rates proportional to the value per kilogram of the commodity was a primitive but effective way of discriminating according to the inverse of the elasticity of demand.

The systematic study of transport demand by professional economists and statisticians began after World War II. The first studies were generally statistical analyses of aggregate data, where the emphasis was primarily on modal choice. The distribution of aggregate freight traffic modes was analysed by regression analysis. It was, however, difficult to interpret the results in the usual form of elasticities of demand or substitution, since the functional forms and theoretical specification had no obvious basis in economic theory of either the firm or of the consumer. Moreover, the levels of aggregation tended to be too large to be of use in deriving useful estimates of parameters. The thrust of much research in the 1970s and 1980s changed to disaggregated data and models. The functional forms were derived from the traditional theory of the consumer (for passenger travel) and from the cost-minimizing motive of the theory of the firm (for freight transportation).

For passenger transport, perhaps the most important econometric development was that of the logit and probit models (Winston 1985). This models the discrete choice of transportation mode in terms of maximizing utility, attributing to each individual a random element of utility which describes his peculiarities and unobservable tastes for travel. Thus, for any given measured characteristics of the population and for given conditions of transportation, one will observe a certain

fraction using a particular model. This probability of using that mode is then described as a functional form of the characteristics of the modes and, of course, other social and economic characteristics of the traveller.

Essentially these models give decision rules for discrete choices. In another sense they provide a statistical description of our ignorance, by specifying a probability density distribution to delineate what we do *not* know about the choice of mode. One objective is to try to reduce this degree of ignorance by defining groupings, such as commuters and shoppers, which have distinctly different patterns of behaviour. A second objective is the joint modelling both of the decision on which mode to use and the choice of another continuous variable, such as frequency of trip or (in the case of freight transport) the size of the consignment or shipment. The continuous choice decision is then contingent upon the discrete decision on mode. This opens up a rich seam of possible models.

Notwithstanding the statistical complexities of these models, there are some basic economic constraints on transport demand. First it is a *derived* demand and so must conform to the Marshallian laws. If there is no conceivable substitution of different modes or markets, then the market elasticity of demand for transport is measured by the multiple of the final market elasticity of demand for the commodity and the fraction of the final price accounted for by transport costs. On such assumptions elasticities of demand are low, even very low. The transport of such items as lumber, minerals and even coal – especially by ocean – approximates such conditions. Studies have confirmed that the freight–tariff elasticities are indeed very low – less than 0.1 (in absolute terms). Yet very few transport operations conform closely to these assumptions. Even in the case of minerals, coal or timber, there is often an alternative source which topples the demand curve from its near vertical slope. Since alternative sources probably become available in discrete jumps, it may well be that the elasticity is very low within a small range, but as the alternative becomes profitable, the elasticity rises substantially. Thus one finds that historically, as transport costs have declines in real terms, the near monopoly of local sources has been eroded as more distant sources become profitable to tap. For freight operations where other *modes* are competitive, the market elasticity of demand for any given mode tends to be very high (absolutely) because of the ease of substitution, for example of truck for rail, or air for truck.

Even in freight transport it seems that transit time is of considerable importance in defining the quality of the service. Speeding the transit and, perhaps even more important, improving its reliability, are attributes which command high premiums for many goods. The shipper saves on inventory costs and maintains a tighter control of the production and distribution process. The numerical values of freight elasticities of demand with respect to time of transit tend to be of the order of 0.4–0.7 for all commodities except for bulk-hauled minerals. The more speedy transit by truck has explained much of the switch of traffic from slow rail to fast truck in the past half century.

The time-of-trip and, *a fortiori*, the waiting time are even more important in explaining the modal choice of passengers, particularly in city transportation. The time-of-trip elasticity tends to be larger (absolutely) than the fare elasticity. But for most studies of passenger travel the market elasticities tend to be less than one (in absolute terms) for a particular model. The main exceptions are for leisure travel by rail, bus and air, where the elasticities tend to be around two. This is consistent with the profitability of the policy of some railways in promoting cheap 'excursion fares' as a way of segmenting the market. Of course, individual airlines or buses, in a competitive environment, will always find that the elasticities with respect to both price and time will be considerably larger than the corresponding market elasticities. So an individual operator will have the greater incentive to keep his fares low and beat the competition. The relatively low elasticity of demand for the mode, compared with the high elasticities for competitive firms within a particular mode (excepting perhaps the railroads), illustrates the value of ensuring a competitive discipline on both pricing and quality of service characteristics.

The high value that people place on time and convenience has many other implications for the structure of transportation services. For example, it may be much better to have more frequent (if more costly) services than those which are produced by choosing vehicles and equipment that minimize the costs to the operator. The competitive operators might well find that the savings in pecuniary cost by adopting the large vehicle with substantial economies of scale are more than offset by the reduction in demand for their services caused by the lower frequency of service – and so longer waiting times. The higher the levels of income of the passengers, the greater the value that they will place on high frequencies and greater convenience.

Investment

The main use of the studies of production and cost functions, on the one hand, and the demand functions, on the other, is to develop a critique of pricing policy and to provide a basis for investment appraisal. The basic criterion is to find the net present value (NPV) of future revenues and outgoings by discounting with a rate of interest that measures the marginal productivity of capital in alternative uses. The rule is that the investment is worthwhile if the NPV exceeds zero. There are many alternative criteria which have been suggested to supplement or supplant the NPV principle, such as the benefit–cost ratio, the first-year rate of return and the implicit rate of return. Although the latter is useful as an evocative and easily understood 'rate of profit', none should be allowed to supplant the NPV method, since all may involve substantial error in application.

The benefits and costs to be entered into the calculation normally assume that the price of the service is, under competitive market conditions, an appropriate indication of the value that the consumer places on the marginal unit that he buys, and that the supply price is competitively determined and reflects marginal costs. These general conditions seem to be widely satisfied in the road sector, and in rather more qualified ways in rail, airlines and ocean or inland waterway transport.

Many of the complications of investment appraisal arise, as in the cost studies, from the multiproduct nature of the services. An improvement of a particular road may displace traffic or generate new vehicle flows on many another facility, and it is often very important to take into account such secondary effects, such as the benefits which consumers may lose (or gain) through the additional congestion (or decongestion) of existing highways. The procedure recommended by theory is to measure the areas under the demand curves for individual roads, while holding the cost conditions on all other roads fixed. This tedious process can be circumvented if one is willing to assume, first, that all demand curves are compensated, and second, that we may approximate the demand curve for the services of the road by a linear form over the relevant range of traffic. Then the gain to road users is

$$-0.5\sum_i (q_{1i} - q_{0i})(p_{1i} - p_{0i})$$

where q_{0i} is the number of i-type trips without the project, q_{1i} is the number of i-type trips with the project and p_{0i} and p_{1i} are the associated prices of trips.

This procedure is much simpler than the traditional Marshallian method of moving one price at a time, and it works well in practical situations. Yet it is important to recall that it depends crucially on the efficacy of the modelling system for road traffic and on the various valuations of cost, particularly the cost of time.

It is relatively easy to estimate the benefits of the increase in efficiency of the use of assets, such as vehicles or runways or terminal capacity. One can estimate the expenditure so avoided by making verifiable assumptions about the rates of utilization. In practice most investment appraisals involving passengers turn critically on the modelling of time savings and the valuation of such savings. Occasionally problems of valuation are of dominant importance in investments in freight transport (such as the widening of the Suez Canal) when there is expected to be a large and persistent

overhang of excess capacity for some years; but such cases are the exception rather than the rule.

By the late 1960s most OECD governments had developed systematic procedures for the appraisal of government investments in transport. Such an approach enabled governments more readily to identify white elephants and to sequence their programme in a more efficient way. However, it is not at all clear that the improved appraisals had any substantial effect on programmes, at least until the austerities of the mid-1970s. For example, the interstate highway system in the United States saw considerable overinvestment in roads that carried little traffic. The political pressures for particular investments were often of much more importance in the decision-making process than any rational calculations of costs and benefits (Walters 1978). The list of undesirable and loss-making projects is depressingly long and formidable for most countries. Apart from such prominent cases as Concorde, one must include perhaps most railway and subway investment, many airports, and some ports on such a list.

The Political Economy of Transport – Inequity and Inefficiency

Most transport activities normally involve considerable state intervention, not merely state participation but often actual ownership of transport concerns. Hence the problem of understanding transportation industries, the consequences of exogenous change and the effects of policies need to be studied in terms of the political process and the power of interest coalitions. Many democratic legislatures have arrogated to themselves the power to control closely the services, rates and fares, investment and manning programmes of state-owned or regulated industries. This power is most obviously demonstrated in the inability to reduce unprofitable rail, bus or airline services. Coalitions of legislature members will form to block withdrawal of service or closure of the line, even though the benefits derived by the constituents comprise only a small fraction of the total cost involved in providing the service. Although it would be more efficient to 'buy off' the users of the service, a sense of misplaced propriety combined with legal constraint has inhibited the development of such side-payments.

Because of the concentration of ownership – usually in the state itself – railways and ports tend to be fertile ground for the formation of powerful trades unions. By exercising control over the arteries of commerce, the unions can exact much danegled. Perhaps more important is that their power is deployed in legislatures and executive branches of government. Such power can be quite awesome – as for example that which the railway trade union exercised in Argentina in the 1950s and 1960s.

Although the problem of exercising control and financial discipline over state-supported or state-owned industries has been seen to be of crucial importance – at least since the end of World War II – little progress appears to have been made. The (Herbert) Morrison formula, used in the UK, was to allow the nationalized railways to operate without political interference in day-to-day decision-making, the legislature and executive having control only over broad policy issues. And the public corporation was to be required to cover its costs, 'taking one year with another'. The failure of the Morrison approach was complete. Not only was there continuous political interference and substantial and persistent large losses, political control over the strategy was, until the 1980s, largely dominated by the trades unions. Conditions in other OECD countries, although varying in detail, are quite similar. Even in the Third World countries, such as India, the power of the rail unions (together with the bureaucrats of Indian Railways) has been sufficient to keep competitive road operators under very large handicaps, and to maintain rail operations when they should long have been superseded by road.

The solutions to the overweening power of nationalized transport organizations and their trades unions, and the politicization of transport decisions, have not come easily. Perhaps the best approach is the privatization of such large bureaucratic organizations. The United States with its dispersion of ownership of railroads seems to

have avoided many of the problems that affect other OECD countries with their nationalized concerns. (In the United States the exception has been Amtrak, the subsidized passenger service.) But, once nationalized, it is not easy to disentangle the integrated system and dispose of its parts. This approach has proved much more successful in the case of the ports (perhaps in the latter half of the 1980s, airports as well) airlines and even buses (World Bank 1985). Various other approaches have been proposed, such as a constitutional limitation on the legislature's directives combined with suitable incentive payments for staff and management; but such solutions have yet to be tried. The inefficiencies of state-owned and regulated systems remain the major problem.

Productivity, Technology and the Development of Transport

Over the historical record the measures of the productivity of many transport industries have shown a creditable performance, relative to most other industries. The only marked exceptions are the railways and urban buses. For highway transport, airlines and certain shipping services, the gains in productivity have exceeded those in most manufacturing industries. In large part such gains have been due to the improvement of equipment used by the transport industries. The increase in the efficiency of aircraft, ships and road vehicles have been both sustained and, in some cases, dramatic. In the decades from 1950 to 1980 one of the main gains in efficiency was due to the increased capacity of equipment – the jumbo tankers and aircraft. The variety and quality of service, however, have also expanded considerably.

One of the most important innovations was a product of the transport industries as such, namely, containerization and its forerunner, palletization. This simplified the handling problems for freight (particularly for non-bulk, non-mineral consignment), greatly reduced inputs of unskilled labour, speeded up transit and reduced pilferage and damage. Containerization, having transformed ports, shipping and much trucking, has still (in 1985) much more potential penetration ahead, with profound effects on distribution systems.

The silicon chip must affect the future of transport. The use of the developing information technology is still very much in its infancy in the transport sector. The potential is great. For example, a computerized road-pricing system for urban areas is now feasible, and prototypes are being tested (World Bank 1985), while plausible control and guidance systems have gone beyond the drawing-board stage. But the range of profitable application cannot be foreseen.

Most of the gain in future productivity and reductions of cost may come from institutional change. It has been shown that privatization of public sector transport operations often has a dramatic effect in reducing costs and improving the quality of the service (World Bank 1985). Private provision of urban bus services usually costs about 50 and 60% of the costs of public provision, and the private airlines in the United States provide services which are only about 40% of the (adjusted) costs of the state-owned airlines in Europe. In addition, the record of the private sector in promoting and paying for innovations is far better than that of the public sector firms. However, whether such innovations will persist in the face of the opposition of vested interests is a question which must remain unanswered.

See Also

- ▶ Congestion
- ▶ Derived Demand
- ▶ Industrial Organization
- ▶ Regulation and Deregulation

Bibliography

Caves, D., L. Christensen, and J. Swanson. 1981, December. Productivity, growth, scale economies and capacity utilization in US railroads, 1955–1974, *American Economic Review* 71(5): 994–1002.

Friedlander, A.F. 1981. *Freight transport regulation*. Cambridge, MA: MIT Press.

Hotelling, H. 1938. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6(July): 242–269.
Keeler, T.E. 1983. *Railroads, freight and public policy.* Washington, DC: Brookings Institution.
Lardner, D. 1850. *Railway economy: A treatise on the new art of transport*. New York: Harper & Bros (Reprinted, New York: A.M. Kelley, 1968).
Meyer, J., M.J. Peck, J. Stenason, and C. Zwick. 1959. *The economics of competition in the transportation industries*. Cambridge, MA: Harvard University Press.
Walker, G. 1948. *Road and rail*. London: Allen & Unwin.
Walters, A.A. 1968. *The economics of road user charges*. Baltimore: Johns Hopkins Press.
Walters, A.A. 1978. Airports – An economic survey. *Journal of Transport Economics and Policy* 12(May): 125–160.
Winston, C. 1985. Conceptual developments in the economics of transportation: an interpretive survey. *Journal of Economic Literature* 23(March): 57–94.
Wohl, M., and C. Hendrickson. 1984. *Transportation investment and pricing principles: An introduction for engineers, planners and economists*. New York: John Wiley & Sons.
World Bank. 1985. *Urban transport sector policy paper.* Washington, DC.

Transversality Condition

Robert A. Becker

Abstract

The transversality condition for an infinite horizon dynamic optimization problem is the boundary condition determining a solution to the problem's first-order conditions together with the initial condition. The transversality condition requires the present value of the state variables to converge to zero as the planning horizon recedes towards infinity. The first-order and transversality conditions are sufficient to identify an optimum in a concave optimization problem. Given an optimal path, the necessity of the transversality condition reflects the impossibility of finding an alternative feasible path for which each state variable deviates from the optimum at each time and increases discounted utility.

Keywords

Arbitrage; Asset pricing models; Bubbles; Capital accumulation programmes; Competitive equilibrium; Depreciation; Euler equations; Infinite horizons; Optimal growth paths; Ramsey model; Transversality condition

JEL Classifications

D4; D10

The transversality condition for an infinite horizon dynamic optimization problem acts as the boundary condition determining a solution to the problem's first-order conditions together with the initial condition. Malinvaud (1953) introduced the transversality condition as part of the sufficient conditions for intertemporally efficient capital accumulation programmes. He required the present value of the capital stock to converge to zero as the planning horizon tended towards infinity. An *efficient programme* is a feasible path of capital stocks starting from a given initial stock, together with a consumption path, having the property that no other feasible programme from the same starting stock provides as much consumption in every period and more consumption in at least one period. He showed that it was possible for a programme to be efficient for any finite planning horizon yet be inefficient when the program was considered over the entire infinite horizon. These inefficient programmes 'over-accumulated' capital and failed to satisfy the transversality condition.

Current theories emphasize the transversality condition's necessity. This is illustrated for the discrete time one-sector discounted Ramsey optimal growth model. There is a single all purpose consumption good, c_t, produced using capital goods, k_{t-1}, carried over from the previous period, and fixed labour. The planner decides how much to consume in the current period and how much to save for next period's production. Capital depreciates entirely within the period. The planner's initial stock of capital produces goods available in the first period. The planner obtains utility, $u(c_t)$, from consumption at time t and maximizes

the discounted sum of future utilities. The discount factor, δ, is a given constant.

The planner's problem is:

$$\sup \sum_{t=1}^{\infty} \delta^{t-1} u(c_t) \text{ by choice of } \{c_t, k_{t-1}\}_{t=1}^{\infty}, \quad (1)$$

subject to :

$$c_t + k_t \leq f(k_{t-1}) \text{ for } t = 1, 2, \ldots;$$
$$c_t \geq 0, k_{t-1} \geq 0 \text{ all } t; k_0 \leq k, \text{ where } k > 0 \text{ is given.} \quad (2)$$

Feasible programs are sequences $\{c_t, k_{t-1}\}_{t=1}^{\infty}$ which satisfy (2). Assume $u : [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $u(0) = 0$, and satisfies the Inada condition: $\lim_{c \to 0+} u'(c) = \infty$. The production function $f : [0, \infty) \to [0, \infty)$ is strictly concave, increasing, twice continuously differentiable, $f(0) = 0$, satisfies $\lim_{k \to 0+} f'(k) = \infty$, and $\lim_{k \to 0+} f'(k) < 1$. There is a maximum sustainable stock, $b > 0$, with $f(b) = b$ and $0 < k < b$. The discount factor satisfies $0 < \delta < 1$. There is a unique optimal program, $\{\overline{c}_t, \overline{k}_{t-1}\}_{t=1}^{\infty}$.

The optimal program satisfies $(\overline{c}_t, \overline{k}_{t-1}) > 0$ for each t. The Kuhn–Tucker necessary conditions for an optimum, known as the *Euler*, or *no-arbitrage conditions*, are:

$$\delta f'(\overline{k}_t) u'(\overline{c}_{t-1}) = u'(\overline{c}_t), \quad \text{for each } t. \quad (3)$$

If the planner's horizon is a finite period, T, then (3) and the complementary slackness condition $\delta^{T-1} u'(\overline{c}_T) \overline{k}_T = 0$ obtain. The latter condition states capital's terminal value is zero. For the infinite horizon case of interest, it is natural to conjecture the *transversality* condition holds as a necessary condition for optimality:

$$\lim_{T \to \infty} \delta^{T-1} u'(\overline{c}_T) \overline{k}_T = 0. \quad (4)$$

The optimal path of the infinite horizon problem converges monotonically to the stationary optimal programme (c^*, k^*), with $c^* = f(k^*) - k^*$ and $\delta f'(k^*) = 1$. Hence, the transversality condition holds. The economic intuition underlying the necessity of the transversality condition is independent of this convergence result.

Equation (3) expresses the unprofitability of the one-period reversed arbitrages developed below. An *arbitrage* represents a feasible change in the optimal path. *Reversed arbitrages* perturb the optimum for finitely many consecutive periods. *Unreversed arbitrages* change the optimal path permanently from some given time on to infinity. A necessary condition for an optimal path is that no arbitrage increase the discounted sum of future utilities above the optimal discounted utility. The necessity of the transversality condition can be interpreted as a type of no-arbitrage condition for unreversed arbitrages.

Assume $\{\overline{c}_t, \overline{k}_{t-1}\}_{t=1}^{\infty}$ is optimal. Suppose the planner decides to increase the first period's consumption and forgoes one unit of capital to be used in next period's production. The marginal gain is $u'(\overline{c}_1)$ in units of utility at time 1. A *T-period reversed arbitrage* occurs if at time $T + 1$ the planner reacquires the unit of capital forgone at time one. After time $T + 1$, the arbitrage no longer affects the path.

Two costs are incurred by the acquisition at time $T + 1$. First, there is the *direct cost* or *repurchase cost* of forgone consumption, which arises from converting a unit of consumption at time $T + 1$ to a unit of capital to be saved for the next period's production. This direct cost equals $\delta^T u'(\overline{c}_{T+1})$ in period 1 utils. The indirect cost arises because the net marginal product of that unit of capital is lost to the planner in every period between $t = 2$ and $t = T + 1$. The indirect cost at time t in utils of time 1 is $\delta^{t-1} u'(\overline{c}_t)[f'(\overline{k}_{t-1}) - 1]$; adding over $t = 2; , \ldots ; T + 1$ yields the present value (focal date one) of the indirect cost. The total cost equals the sum of the direct and indirect costs:

$$\sum_{t=2}^{T+1} \delta^{t-1} u'(\overline{c}_t)[f'(\overline{k}_{t-1}) - 1] + \delta^T u'(\overline{c}_{T+1}). \quad (5)$$

A necessary condition for the optimality is, for any T, the marginal benefit of a T-period reversed arbitrage is equal to its marginal (discounted) cost. Thus,

$$u'(\overline{c}_1) = \sum_{t=2}^{T+1} \delta^{t-1} u'(\overline{c}_t)\left[f'\left(\overline{k}_{t-1}\right) - 1\right] + \delta^T u'(\overline{c}_{T+1}). \quad (6)$$

Equation (6) applied to a one-period reversed arbitrage reduces to (3), evaluated at $T = 1$. Equation (3) follows from the same reasoning when a reversed arbitrage starts at time t.

Equation (6) contains no further information. However, the infinite horizon means the planner can also contemplate the profitability of an *unreversed* arbitrage in which the unit of capital is permanently sacrificed at time $t = 1$. There is no repurchase cost associated with an unreversed arbitrage, hence the conditions for the unprofitability of an unreversed arbitrage must be

$$u'(\overline{c}_1) = \sum_{t=2}^{\infty} \delta^{t-1} u'(\overline{c}_t)\left[f'\left(\overline{k}_{t-1}\right) - 1\right]. \quad (7)$$

But (6) and (7) can hold as $T \to \infty$ only if $\lim_{T\to\infty} \delta^T u'(\overline{c}_{T+1}) = 0$, which implies the transversality condition since the capital stocks are bounded. An unreversed arbitrage over-accumulates capital in comparison with the optimal programme. This cannot be optimal. Thus, the transversality condition expresses the zero marginal profit condition for the open-ended arbitrages which are available only in the infinite horizon context. The rate of capital accumulation is thereby limited.

The Euler equations and the transversality conditions are necessary for optimality. For concave optimal growth models – the case where utility and production functions are concave – the Euler equations and transversality conditions are also sufficient to identify an optimal programme. Indeed, transversality conditions are necessary (and sufficient) together with the Euler equations in other dynamic models where the state variables need not be capital stocks.

The transversality condition's necessity is important in connecting dynamic equilibrium paths and optimal growth paths in infinitely lived representative agent economies. The basic idea is to match the Euler and transversality conditions in the equilibrium and optimal growth settings. This results in an equivalence principle: an optimal growth path is a competitive equilibrium, and vice versa.

The necessity of the transversality condition can also be used to exclude bubbles from occurring in asset pricing models. The asset's fundamental is the present discounted value of its future dividend payouts. A bubble exists if the asset's price differs from its fundamental value. For example, if the asset is a perpetuity and offers a constant dividend stream, then the no-arbitrage conditions state the capital gain yield plus the dividend yield equals the interest rate at each time. Formally,

$$\frac{p_{t+1} - p_t}{p_t} + \frac{r}{p_t} = \varrho, \quad t = 1, 2, \ldots. \quad (8)$$

where p_t is the asset's current price at time t, r is the asset's return (or dividend) in each period, and R is the constant interest rate.

Equation (8) is a first-order difference equation. Its solution is

$$p_t = (1+\varrho)^{t-1}\left[p_1 - \left(\frac{r}{\varrho}\right)\right] + \left(\frac{r}{\varrho}\right). \quad (9)$$

For each choice of p_1, there is a sequence, $\{p_t\}_{t=1}^{\infty}$, calculated from (9). Thus, there are an infinite number of price systems satisfying this asset's no-arbitrage equation.

Efficient markets would imply the absence of arbitrage opportunities for all time and single out the solution $p_t = (r/\varrho)$, which occurs if and only if $p_1 = (r/\varrho)$. Initial prices with $p_1 > (r/\varrho)$ would create a bubble where p_t exceeds its *fundamental value*, (r/ϱ). Prices continue to rise simply because investors expect them to do so. There is a *negative bubble* if $p_1 < (r/\varrho)$. The asset's price becomes negative in finite time, an impossibility as prices must always remain non-negative. Hence, $p_1 \geq (r/\varrho)$.

The transversality condition takes the form

$$\lim_{t\to\infty} p_t \left(\frac{1}{1+\varrho}\right)^{t-1} = 0. \quad (10)$$

If this is an equilibrium condition, then the initial price must be $p_1 = (r/\varrho)$ and the asset's market price equals its fundamental at each time.

Imposition of the transversality condition picks out the only solution of (8) that is not a bubble.

Many deterministic models have stochastic counterparts. For example, technology shocks in the Ramsey problem lead to stochastic Euler equations expressing no-arbitrage opportunities in expectations, or on average, when the planner's objective is the expected discounted sum of future utilities. The corresponding transversality condition also holds in expectations. There can exist particular realizations of the shocks for which a bubble persists. However, on average, there are no unprofitable unreversed arbitrages.

The argument for the necessity of the transversality condition given above is heuristic. Weitzman (2003) presents an analogous intuitive rationale for the transversality condition in continuous time. The terms 'reversed' and 'unreversed' arbitrage originate in Gray and Salant (1983). Rigorous arguments are found in the references below.

See Also

- Arbitrage Pricing Theory
- Bubbles
- Efficient Markets Hypothesis
- Government Budget Constraint
- Tulipmania

Bibliography

Becker, R.A., and J.H. Boyd. 1997. *Capital theory, equilibrium analysis, and recursive utility.* Malden: Blackwell Publishers.

Benveniste, L.M., and J.A. Schienkman. 1982. Duality theory for dynamic optimization models of economics: The continuous time case. *Journal of Economic Theory* 27: 1–19.

Gray, J.A. and S.W. Salant 1983. Transversality conditions in infinite horizon models. Working paper, Washington State University.

Kamihigashi, T. 2005. Necessity of the transversality condition for stochastic models with bounded or CRRA utility. *Journal of Economic Dynamics and Control* 29: 1313–1329.

LeRoy, S.F. 2004. Rational exuberance. *Journal of Economic Literature* 41: 783–804.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21: 233–268.

Michel, P. 1982. On the transversality condition in infinite horizon optimal problems. *Econometrica* 50: 975–986.

Mirman, L.J., and I. Zilcha. 1975. Optimal growth under uncertainty. *Journal of Economic Theory* 11: 329–339.

Weitzman, M.L. 1973. Duality theory for infinite horizon convex models. *Management Science* 19: 783–789.

Weitzman, M.L. 2003. *Income, wealth, and the maximum principle*. Cambridge, MA: Harvard University Press.

Transversality Conditions and Dynamic Economic Behaviour

Takashi Kamihigashi

Abstract

Transversality conditions are optimality conditions often used along with Euler equations to characterize the optimal paths of dynamic economic models. This article illustrates the role of transversality conditions in characterizing optimal paths as well as in ruling out economic phenomena such as asset bubbles and hyperdeflations in infinite-horizon models.

Keywords

Bubbles; Calculus of variations; Dynamic models; Dynamic optimization; Euler equations; Hyperdeflation; Infinite horizons; Optimality; Overlapping generations models; Ponzi games; Ramsey model; Transversality condition; Transversality conditions and dynamic economic behaviour

JEL Classifications

C61; E1

Transversality conditions are optimality conditions often used along with Euler equations to characterize the optimal paths of dynamic economic models. The purpose of this article is to illustrate the role of transversality conditions in characterizing optimal paths as well as in ruling out economic phenomena such as asset bubbles and hyperdeflations in infinite-horizon models.

See transversality condition for mathematical foundations.

A Geometric Example

A simple geometric example best illustrates the mathematical roles of an Euler equation and a transversality condition. What is the shortest path from a point A to a straight line L infinitely long in both directions? The answer is, of course, the straight line from point A to line L that is perpendicular to line L. There are two conditions involved here. The first condition is that the shortest path be a straight line: one cannot make the path shorter by deviating from it and eventually returning to it. This is the implication of the Euler equation for this problem. But there are infinitely many straight lines from point A to line L. In fact, a straight line from point A to line L can be arbitrarily long; even very bad choices satisfy the Euler equation. This is why one needs the second condition, that the shortest path be perpendicular to line L. This additional condition ensures that one cannot make the path shorter by deviating from it and never returning to it.

The condition of perpendicularity in this example and similar conditions on end points in other problems are called transversality conditions in dynamic optimization theory (Hestenes 1966, p. 87). According to Bolza (1904, p. 106), the term was first introduced by Kneser (1900). Both Euler equations and transversality conditions were initially developed for calculus-of-variations problems. In economics, in particular in macroeconomics, both types of conditions are used mainly for infinite-horizon models in discrete time as well as in continuous time. In what follows, we focus on discrete-time models, which are technically easier to deal with.

A General Discrete-Time Problem

Consider the following maximization problem:

$$\max_{\{x_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t v(x_t, x_{t+1}) \tag{1}$$

$$\text{s.t.} \quad x_{t+1} \geq 0, \quad t = 0, \ 1, \ 2, \ldots, \tag{2}$$

$$x_0 \quad \text{given}, \tag{3}$$

where $\beta \in (0,1)$ is called the discount factor, v is called the return function, and x_t is an n-dimensional vector. There may be other constraints, but we assume that they are never binding at the optimum. We also assume that the nonnegativity constraint is never binding at the optimum, and that the return function v is differentiable and concave. Though the return function may depend on t, we do not assume so here for notational simplicity. To be concrete, we interpret x_t as the stock of wealth (or capital) at the beginning of period t. In most economic problems, it is costly to accumulate wealth. Hence we assume that $v_2(x, y) \leq 0$ for all x, y.

The Euler equation for this problem is simply the first-order condition with respect to x_{t+1}:

$$v_2(x_t, x_{t+1}) + \beta v_1(x_{t+1}, x_{t+2}) = 0. \tag{4}$$

This condition means that no gain can be achieved by deviating from an optimal path for one period. The transversality condition is given by

$$\lim_{T \to \infty} \beta^T \left[-v^2(x_T, x_{T+1})\right] x_{T+1} = 0. \tag{5}$$

Though this is the typical transversality condition in economics, other transversality conditions are possible depending on the specification of constraints. Condition (5) can be interpreted as saying that the present discounted value of wealth at infinity must be zero, or wealth (x_{T+1}) should not grow too fast compared with its discounted marginal value ($\beta^T[-v_2(x_T, x_{T+1})]$). In other words, the transversality condition (5) rules out overaccumulation of wealth. The idea is that, if one saves too much and spends too little, then one is not behaving optimally.

It is well known that the Euler equation (4) and the transversality condition (5) are sufficient for optimality (Stokey and Lucas 1989, p. 89). This result is often credited to Mangasarian (1966), who showed a finite-horizon version of the result.

Since the Euler equation is simply the first-order condition with respect to x_{t+1}, it is a necessary condition for optimality. On the other hand, necessity of the transversality condition is often considered to be a difficult issue. But there are two simple ways to prove its necessity if the objective function is assumed to be finite for all feasible paths (Kamihigashi 2002, 2005). If this assumption fails, one can try the following test. Shift the entire optimal path downward by a small fixed proportion. Does it reduce the value of the objective function by only a finite amount? If so, the transversality condition is necessary. See Kamihigashi (2001, 2003) for precise assumptions and statements. See Weitzman (1973), Benveniste and Scheinkman (1982), and Michele (1982) for earlier results and arguments.

The Ramsey Model

To see how the transversality condition can be used in practice, consider the basic Ramsey model:

$$\max_{\{c_t, x_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{6}$$

$$\text{s.t.} \quad c_t + x_{t+1} = f(x_t), \quad c_t, x_{t+1} \geq 0, \quad t = 0, \ 1, \ 2, \ldots, \tag{7}$$

$$x_0 \quad \text{given}, \tag{8}$$

where c_t is consumption and x_t is the stock of capital at the beginning of period t. We assume that the utility function u and the production function f are continuously differentiable, strictly increasing, and strictly concave. We also assume that $f(0) = 0$, $\lim_{c \to 0} u'(c) = \lim_{x \to 0} f'(x) = \infty$, and $\lim_{x \to \infty} f'(x) < 1$. The model here is a special case of the general problem described above with $v(x_t, x_{t+1}) = u(f(x_t) - x_{t+1})$. The Euler equation and the transversality condition are

$$u'(c_t) = \beta u'(c_{t+1}) f'(x_{t+1}), \tag{9}$$

$$\lim_{T \to \infty} \beta^T u'(c_T) x_{T+1} = 0. \tag{10}$$

Given any initial capital stock x_0, there are three types of paths obeying (7) and (9). Specifically, there is a unique level of consumption $\hat{c}_0$ such that a path $\{c_t, x_{t+1}\}$ satisfying (7) and (9) converges to the unique steady state (which is determined by the strictly positive capital stock x^* such that $\beta f'(x^*) = 1$) if and only if $c_0 = \hat{c}_0$. If $c_0 > \hat{c}_0$, then (7) is eventually violated; such paths are ruled out on feasibility grounds. If $c_0 < \hat{c}_0$, then c_t converges to zero and x_t converges to the capital stock $\bar{x}$ given by $f(\bar{x}) = \bar{x}$. It can be shown that such paths violate the transversality condition and thus are not optimal. The path converging to the steady state satisfies the transversality condition since c_t and x_t converge to their steady state values; hence this is the optimal path.

The preceding argument shows that when one restricts attention to the dynamical system defined by (7) and (9), most paths do not converge to the steady state. This is an example of the Hahn problem (Hahn 1987). The Hahn problem disappears here when one takes the transversality condition into account, since only the path converging to the steady state satisfies the transversality condition as well as the feasibility requirements.

Asset Bubbles

Transversality conditions are often used to rule out asset bubbles. To be specific, consider a deterministic version of the Lucas (1978) asset pricing model. There are many homogeneous agents, a single good, and a single asset that pays a dividend of d_t units of the good in each period t. The population of agents is normalized to one; so is the supply of the asset. Each agent solves

$$\max_{\{c_t, x_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{11}$$

$$\text{s.t.} \quad c_t + p_t x_{t+1} = (p_t + d_t) x_t, \quad c_t, x_{t+1} \geq 0, \quad t = 0, \ 1, \ 2, \ldots, \tag{12}$$

$$x_0 = 1 \quad \text{given}, \tag{13}$$

where c_t is consumption, p_t is the price of the asset, and x_t is shares in the asset at the beginning of period t. In equilibrium, $c_t = d_t$ and $x_t = 1$. We assume that the utility function u is concave, differentiable, and strictly increasing. The Euler equation and the transversality condition in equilibrium are

$$u'(d_t)p_t = \beta u'(d_{t+1})\left(p_{t+1} + d_{t+1}\right), \tag{14}$$

$$\lim_{T\to\infty} \beta^T u'(d_T)p_T = 0. \tag{15}$$

It is easy to see that the sequence $\left\{p_t^*\right\}$ given by

$$p_t^* = \sum_{i=1}^{\infty} \beta^i \frac{u'(d_{t+i})}{u'(d_t)} d_{t+i} \tag{16}$$

satisfies the Euler equation (14). The right-hand side of (16) is called the fundamental value of the asset. Let $\{b_t\}$ be any nonnegative sequence satisfying

$$u'(d_t)b_t = \beta u'(d_{t+1})b_{t+1}. \tag{17}$$

Then the sequence $\left\{p_t^* + b_t\right\}$ also satisfies the Euler equation. Hence there are infinitely many paths satisfying the Euler equation. The extra component b_t, which grows at a gross rate of $u'(d_t)/[\beta u'(d_{t+1})]$, is interpreted as a bubble.

Notice that the bubble component b_t, if strictly positive, violates the transversality condition (15) (with $p_T = b_T$). Therefore, if the transversality condition is necessary for optimality, the bubble component must vanish, so that the price must always be equal to the fundamental value. This is indeed the case in this model (Kamihigashi 2001, p. 1007).

In stochastic models, bubbles can be ruled out by the same argument under standard assumptions, but there are pathological cases in which bubbles are possible (Kamihigashi 1998; Montrucchio and Privileggi 2001). Bubbles arise more easily in models with heterogenous agents such as overlapping generations models, where there is no economy-wide transversality condition. See speculative bubbles.

Hyperdeflations

Transversality conditions are often used to rule out hyperdeflations in money-in-the-utility-function models of the type studied by Brock (1974) and Obstfeld and Rogoff (1986). In these models, agents derive utility from real money balances in addition to consumption. As in the Lucas asset pricing model, there are many paths satisfying the Euler equation. Along a path satisfying the Euler equation with a positive bubble, the value of real balances grows unboundedly or, equivalently, the nominal price level keeps declining towards zero. Such paths are often interpreted as exhibiting hyperdeflations. Under reasonable assumptions, hyperdeflationary paths are ruled out by the transversality condition, which once again rules out overaccumulation of wealth, or real balances.

However, there are cases in which the transversality condition does not rule out hyperdeflationary paths (Obstfeld and Rogoff 1986, p. 356). This is because agents, who derive utility from real balances, benefit directly from accumulating wealth.

No-Ponzi-Game Conditions

In formulating a consumer's maximization problem, one must include some constraint on debt, since otherwise the consumer would never pay back his debt, letting it grow unboundedly. One way to rule out this behaviour is to prohibit debt entirely, that is, to require wealth to be always non-negative (as in (12)). A more lenient way is to require only the present discounted value of wealth at infinity to be non-negative. This type of condition is known as a no-Ponzi-game condition (Blanchard and Fischer 1989, p. 49), but often called a transversality condition as well. A no-Ponzi-game condition is a constraint that prevents overaccumulation of debt, while a typical transversality condition is an optimality condition that rules out overaccumulation of wealth. They place opposite restrictions, and should not be confused.

See Also

- ▶ Bubbles
- ▶ Calculus of Variations
- ▶ Euler Equations
- ▶ Ramsey Model
- ▶ Speculative Bubbles
- ▶ Transversality Condition

Bibliography

Benveniste, L.M., and J.A. Scheinkman. 1982. Duality theory for dynamic optimization models of economics: The continuous time case. *Journal of Economic Theory* 27: 1–19.

Blanchard, O.J., and S. Fischer. 1989. *Lectures on macroeconomics*. Cambridge, MA: MIT Press.

Bolza, O. 1904. *Lectures on the calculus of variations*. Chicago: University of Chicago Press.

Brock, W.A. 1974. Money and growth: The case of long run perfect foresight. *International Economic Review* 15: 750–777.

Hahn, F.H. 1987. Hahn problem. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. London: Macmillan.

Hestenes, M.R. 1966. *Calculus of variations and optimal control theory*. New York: Wiley.

Kamihigashi, T. 1998. Uniqueness of asset prices in an exchange economy with unbounded utility. *Economic Theory* 12: 103–122.

Kamihigashi, T. 2001. Necessity of transversality conditions for infinite horizon problems. *Econometrica* 69: 995–1012.

Kamihigashi, T. 2002. A simple proof of the necessity of the transversality condition. *Economic Theory* 20: 427–433.

Kamihigashi, T. 2003. Necessity of transversality conditions for stochastic problems. *Journal of Economic Theory* 109: 140–149.

Kamihigashi, T. 2005. Necessity of the transversality condition for stochastic models with bounded or CRRA utility. *Journal of Economic Dynamics and Control* 29: 1313–1329.

Kneser, A. 1900. *Lehrbuch der Variationsrechnung*. Braunschweig: F. Vieweg und Sohn.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1445.

Mangasarian, O.L. 1966. Sufficient conditions for the optimal control of nonlinear systems. *SIAM Journal of Control* 4: 139–152.

Michele, P. 1982. On the transversality condition in infinite-horizon problems. *Econometrica* 50: 975–985.

Montrucchio, L., and F. Privileggi. 2001. On fragility of bubbles in equilibrium asset pricing models of Lucas-type. *Journal of Economic Theory* 101: 158–188.

Obstfeld, M., and K. Rogoff. 1986. Ruling out divergent speculative bubbles. *Journal of Monetary Economics* 17: 349–362.

Stokey, N., and R.E. Lucas Jr. 1989. *Recursive methods in economic dynamics*. Cambridge, MA: Harvard University Press.

Weitzman, M.L. 1973. Duality theory for infinite horizon convex models. *Management Science* 19: 783–789.

Treatment Effect

Joshua D. Angrist

Abstract

The term 'treatment effect' refers to the causal effect of a binary (0–1) variable on an outcome variable of scientific or policy interest. Economics examples include the effects of government programmes and policies, such as those that subsidize training for disadvantaged workers, and the effects of individual choices like college attendance. The principal econometric problem in the estimation of treatment effects is selection bias, which arises from the fact that treated individuals differ from the non-treated for reasons other than treatment status per se. Treatment effects can be estimated using social experiments, regression models, matching estimators, and instrumental variables.

Keywords

Average treatment effect; Constant-effects models; Identifying assumptions; Instrumental variables (IV) methods; Law of large numbers; Local average treatment effect; Matching estimators; Monotonicity; Omitted variables bias: see selection bias; Potential-outcomes framework; Propensity-score matching; Regression models; Selection bias; Switching regressions model; Treatment effect; Two-stage least squares; Two-step estimators; Wald estimator

JEL Classifications

C14; C21; C31

A 'treatment effect' is the average causal effect of a binary (0–1) variable on an outcome variable of scientific or policy interest. The term 'treatment effect' originates in a medical literature concerned with the causal effects of binary, yes-orno 'treatments', such as an experimental drug or a new surgical procedure. But the term is now used much more generally. The causal effect of a subsidized training programme is probably the mostly widely analysed treatment effect in economics (see, for example, Ashenfelter 1978, for one of the first examples, or Heckman and Robb 1985 for an early survey). Given a data-set describing the labour market circumstances of trainees and a non-trainee comparison group, we can compare the earnings of those who did participate in the programme and those who did not. Any empirical study of treatment effects would typically start with such simple comparisons. We might also use regression methods or matching to control for demographic or background characteristics.

In practice, simple comparisons or even regression-adjusted comparisons may provide misleading estimates of causal effects. For example, participants in subsidized training programmes are often observed to earn less than ostensibly comparable controls, even after adjusting for observed differences (see, for example, Ashenfelter and Card 1985). This may reflect some sort of omitted variables bias, that is, a bias arising from unobserved and uncontrolled differences in earnings potential between the two groups being compared. In general, omitted variables bias (also known as selection bias) is the most serious econometric concern that arises in the estimation of treatment effects. The link between omitted variables bias, causality, and treatment effects can be seen most clearly using the potential-outcomes framework.

Causality and Potential Outcomes

The notion of a causal effect can be made more precise using a conceptual framework that postulates a set of potential outcomes that could be observed in alternative states of the world. Originally introduced by statisticians in the 1920s as a way to discuss treatment effects in randomized experiments, the potential outcomes framework has become the conceptual workhouse for non-experimental as well as experimental studies in many fields (see Holland 1986, for a survey and Rubin 1974, 1977, for influential early contributions). Potential outcomes models are essentially the same as the econometric *switching regressions* model (Quandt 1958), though the latter is usually tied to a linear regression framework. Heckman (1976, 1979) developed simple two-step estimators for this model.

Average Causal Effects

Except in the realm of science fiction, where parallel universes are sometimes imagined to be observable, it is impossible to measure causal effects at the individual level. Researchers therefore focus on average causal effects. To make the idea of an average causal effect concrete, suppose again that we are interested in the effects of a training programme on the post-training earnings of trainees. Let Y_{1i} denote the potential earnings of individual i if he were to receive training and let Y_{0i} denote the potential earnings of individual i if not. Denote training status by a dummy variable, D_i. For each individual, we observe $Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$, that is, we observe Y_{1i} for trainees and Y_{0i} for everyone else.

Let $E[\cdot]$ denote the mathematical expectation operator, i.e., the population average of a random variable. For continuous random variables, $E[Y_i] = \int yf(y)dy$, where $f(y)$ is the density of Y_i. By the law of large numbers, sample averages converge to population averages so we can think of $E[\cdot]$ as giving the sample average in very large samples. The two most widely studied average causal effects in the treatment effects context are the average treatment effect (ATE), $E[Y_{1i} - Y_{0i}]$, and the average treatment effect on the treated (ATET), $E[Y_{1i} - Y_{0i}|D_i = 1]$. Note that the ATET can be rewritten

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1].$$

T

This expression highlights the counter-factual nature of a causal effect. The first term is the average earnings in the population of trainees, a potentially observable quantity. The second term is the average earnings of trainees had they not been trained. This cannot be observed, though we may have a control group or econometric modelling strategy that provides a consistent estimate.

Selection Bias and Social Experiments

As noted above, simply comparing those who are and are not treated may provide a misleading estimate of a treatment effect. Since the omitted variables problem is unrelated to sampling variance or statistical inference, but rather concerned with population quantities, it too can be efficiently described by using mathematical expectation notation to denote population averages. The contrast in average outcomes by observed treatment status is

$$E[Y_i|D_i=1]-E[Y_i|D_i=0]=E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=0]\ =E[Y_{1i}-Y_{0i}|D_i=1]+\{E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]\}$$

Thus, the naive contrast can be written as the sum of two components, ATET, plus selection bias due to the fact that the average earnings of non-trainees, $E[Y_{0i}|D_i=0]$, need not be a good stand-in for the earnings of trainees had they not been trained, $E[Y_{0i}|D_i=1]$.

The problem of selection bias motivates the use of random assignment to estimate treatment effects in social experiments. Random assignment ensures that the potential earnings of trainees had they not been trained – an unobservable quantity – are well-represented by the randomly selected control group. Formally, when D_i is randomly assigned, $E[Y_i|D_i=1]-E[Y_i|D_i=0]=E[Y_{1i}-Y_{0i}|D_i=1]=E[Y_{1i}-Y_{0i}]$. Replacing $E[Y_i|D_i=1]$ and $E[Y_i|D_i=0]$ with the corresponding sample analogs provides a consistent estimate of ATE.

Regression and Matching

Although it is increasingly common for randomized trials to be used to estimate treatment effects, most economic research still uses observational data. In the absence of a randomized experiment, researchers rely on a variety of statistical control strategies and/or natural experiments to reduce omitted variables bias. The most commonly used statistical techniques in this context are regression, matching and instrumental variables.

Regression estimates of causal effects can be motivated most easily by postulating a constant-effects model, where $Y_{1i}-Y_{0i}=\alpha$ (a constant). The constant-effects assumption is not strictly necessary for regression to estimate an average causal effect, but it simplifies things to postpone a discussion of this point. More importantly, the only source of omitted-variables bias is assumed to come from a vector of observed covariates, X_i, that may be correlated with D_i. The key assumption that facilitates causal inference in regression models (sometimes called an identifying assumption), is that

$$E[Y_{0i}|X_i,D_i]=X_1'\beta, \tag{1}$$

where β is a vector of regression coefficients. This selection-on-observables assumption has two parts. First, Y_{0i} (and hence Y_{1i}, given the constant-effects assumption) is mean-independent of D_i conditional on X_i. Second, the conditional mean function for Y_{0i} given X_i is linear. Given Eq. (1), it is straightforward to show that

$$E\{Y_i(D_i-R[D_i|X_i])\}/E\{D_i(D_i-R[D_i|X_i])\}=\alpha, \tag{2}$$

where $R[D_i/X_i]$ are the fitted values from a regression of D_i, on X_i. This is the coefficient on D_i, from the population regression of Y_i on D_i, and X_i, (that is, the regression coefficient in an infinite sample). Again, the law of large numbers ensures that sample regression coefficients estimate this population regression coefficient consistently.

Matching is similar to regression in that it is motivated by the assumption that the only source of omitted variables or selection bias is the set of observed covariates, X_i. Unlike regression, however, matching estimates of treatment effects are constructed by matching individuals with the same covariates instead of through a linear

model for the effect of covariates. Instead of (1), the selection-on-observables assumption becomes

$$E\left[Y_{ji}|X_i,D_i\right]=E\left[Y_{ji}|X_i\right],\quad \text{for } j=0,1. \tag{3}$$

This implies

$$\begin{aligned}&E[Y_{1i}-Y_{0i}|D_i=1]\\&=E\{E[Y_{1i}|X_i,D_i=1]-[Y_{0i}|X_i,D_i=1]|D_i=1\}\\&=E\{E[Y_{1i}|X_i,D_i=1]-[Y_{0i}|X_i,D_i=0]|D_i=1\}\end{aligned} \tag{4a}$$

and, likewise,

$$\begin{aligned}&E[Y_{1i}-Y_{0i}]\\&=E\{E[Y_{1i}|X_i,D_i=1]-[Y_{0i}|X_i,D_i=0]\}\end{aligned} \tag{4b}$$

In other words, we can construct ATET or ATE by averaging X-specific treatmentcontrol contrasts, and then reweighting these X-specific contrasts using the distribution of X_i, for the treated (for ATET) or using the marginal distribution of X_i (for ATE). Since these expressions involve observable quantities, it is straightforward to construct consistent estimators from their sample analogs.

The conditional independence assumption that motivates the use of regression and matching is most plausible when researchers have extensive knowledge of the process determining treatment status. An example in this spirit is the Angrist (1998) study of the effect of voluntary military service on the civilian earnings of soldiers after discharge, discussed further below.

Regression and Matching Details

In practice, regression can be understood as a type of weighted matching estimator. If, for example, $E[D_i|X_i]$ is a linear function of X_i, (as it might be if the covariates are all discrete), then it is possible to show that Eq. (2) is equivalent to a matching estimator that weights cell-by-cell treatment-control contrasts by the conditional variance of treatment in each cell (Angrist 1998). This equivalence highlights the fact that the most important econometric issue in a study that relies on selection-on-observables assumptions to identify causal effects is the validity of these conditional independence assumptions, not whether regression or matching is used to implement them.

A computational difficulty that sometimes arises in matching models is how to find good matches for each possible value of the covariates when the covariates take on many values. For example, beginning with Ashenfelter (1978), many studies of the effect of training programmes have shown that trainees typically experience a period of declining earnings before they go into training. Because lagged earnings is both continuous and multidimensional (since more than one period's earnings seem to matter), it may be hard to match trainees and controls with exactly the same pattern of lagged earnings. A possible solution in this case is to match trainees and controls on the *propensity score,* the conditional probability of treatment given covariates. Propensity-score matching relies on the fact that, if conditioning on X_i eliminates selection bias, then so does conditioning on $P[D_i=1|X_i]$, as first noted by Rosenbaum and Rubin (1983). Use of the propensity score reduces the dimensionality of the matching problem since the propensity score is a scalar, though in practice it must still be estimated. See Dehejia and Wahba (1999) for an illustration.

Regression and Matching Example

Between 1989 and 1992, the size of the military declined sharply because of increasing enlistment standards. Policymakers would like to know whether the people – many of them black men – who would have served under the old rules but were unable to enlist under the new rules were hurt by the lost opportunity for service. The Angrist (1998) study attempts to answer this question. The regression and matching assumptions seem plausible in this context because soldiers are selected on the basis of a few well-documented criteria related to age, schooling and test scores and because the control group used in the study also applied to enter the military.

Naive comparisons clearly overestimate the benefit of military service. This can be seen in Table 1, which reports differences-in-means, matching and regression estimates of the effect of voluntary military service on the 1988–91 Social Security-taxable earnings of men who applied to join the military between 1979 and 1982. The matching estimates were constructed from the sample analog of (4a), that is, from covariate-value-specific differences in earnings, weighted to form a single estimate using the distribution of covariates among veterans. The covariates in this case are the age, schooling and test-score variables used to select soldiers from the pool of applicants. Although white veterans earn $1233 more than non-veterans, this difference becomes negative once the adjustment for differences in covariates is made. Similarly, while non-white veterans earn $2449 more than non-veterans, controlling for covariates reduces this to $840.

Table 1 also shows regression estimates of the effect of voluntary military service, controlling for the same covariates used for matching. These are estimates of α_r in the equation

$$Y_i = \sum_X d_{iX}\beta_X + \alpha_r D_i + e_i,$$

where β_X is a regression-effect for $X_i = X$ and α_r is the regression parameter. This corresponds to a saturated model for discrete X_i. The regression estimates are larger than (and significantly different from) the matching estimates. But the regression and matching estimates are not very different economically, both pointing to a small earnings loss for White veterans and a modest gain for Non-whites.

Instrumental Variables Estimates of Treatment Effects

The assumptions required for regression or matching to identify a treatment effect are often implausible. Many of the necessary control variables are typically unmeasured or simply unknown. Instrumental variables (IV) methods solve the problem of missing or unknown controls, much as a randomized trial also obviates the need for regression or matching. To see how this is possible, begin again with a constant effects model without covariates, so $Y_{1i} - Y_{0i} = \alpha$. Also, let $Y_{0i} = \beta + \varepsilon_i$, where $\beta \equiv E[Y_{0i}]$. The potential outcomes model can now be written

$$Y_i = \beta + \alpha D_i + \varepsilon_i, \tag{5}$$

where α is the treatment effect of interest. Because D_i is likely to be correlated with ε_i, regression estimates of Eq. (5) do not estimate α consistently.

Now suppose that in addition to Y_i and D_i there is a third variable, Z_i that is correlated with D_i, but unrelated to Y_i for any other reason. In a constant-effects world, this is equivalent to saying

Treatment Effect, Table 1 Matching and regression estimates of the effects of voluntary military service in the United States

Race	Average earnings in 1988–91 (1)	Differences in means (2)	Matching estimates (3)	Regression estimates (4)	Regression minus matching (5)
Whites	14,537	1233.4	–197.2	–88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11,664	2449.1	839.7	1074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

Notes: Figures are in nominal US dollars. The table shows estimates of the effect of voluntary military service on the 1988–91 Social Security-taxable earnings of men who applied to enter the armed forces during 1979–82. The matching and regression estimates control for applicants' year of birth, education at the time of application, and Armed Forces Qualification Test (AFQT) score. There are 128,968 whites and 175,262 non-whites in the sample. Standard errors are reported in parentheses (*Source:* Adapted from Angrist (1998, Tables II and V))

Y_{0i} and Z_i are independent. It therefore follows that

$$E[\varepsilon_i | Z_i] = 0, \tag{6}$$

a conditional independence restriction on the relation between Z_i and Y_{0i}, instead of between D_i and Y_{0i} as required for regression or matching strategies. The variable Z_i is said to be an IV or just 'an instrument' for the causal effect of D_i on Y_i.

Suppose that Z_i is also a 0–1 variable. Taking expectations of (5) with Z_i switched off and on, we immediately obtain a simple formula for the treatment effect of interest:

$$\frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]} = \alpha. \tag{7}$$

The sample analog of this equation is sometimes called the Wald estimator, since it first appear in a paper by Wald (1940) on errors-in-variables problems. There are other more complicated IV estimators involving continuous, multi-valued, or multiple instruments. For example, with a multi-valued instrument, we might use the sample analog of $\text{Cov}(Z_i, Y_i)/\text{Cov}(D_i, Y_i)$. This simplifies to the Wald estimator when Z_i is 0–1. The Wald estimator captures the main idea behind most IV estimation strategies since more complicated estimators can usually be written as a linear combination of Wald estimators (Angrist 1991).

IV Example

To see how IV works in practice, it helps to use an example, in this case the effect of Vietnam-era military service on the earnings of veterans later in life (Angrist 1990). In the 1960s and early 1970s, young men were at risk of being drafted for military service. Concerns about fairness led to the institution of a draft lottery in 1970 that was used to determine priority for conscription in cohorts of 19-year-olds. A natural instrumental variable for the Vietnam veteran treatment effect is draft-eligibility status, since this was determined by a lottery over birthdays. In particular, in each year from 1970 to 1972, random sequence numbers (RSNs) were randomly assigned to each birth date in cohorts of 19-year-olds. Men with lottery numbers below an eligibility ceiling were eligible for the draft, while men with numbers above the ceiling could not be drafted. In practice, many draft-eligible men were still exempted from service for health or other reasons, while many men who were draft-exempt nevertheless volunteered for service. So veteran status was not completely determined by randomized draft eligibility; eligibility and veteran status are merely correlated.

For white men who were at risk of being drafted in the 1970–71 draft lotteries, draft-eligibility is clearly associated with lower earnings in years after the lottery.

This can be seen in Table 2, which reports the effect of randomized draft-eligibility status on

Treatment Effect, Table 2 Instrumental variables estimates of the effects of military service on US white men born 1950

	Earnings		Veteran status		Wald estimate of veteran effect
Earnings year	Mean	Eligibility effect	Mean	Eligibility effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	–435.8	0.267	0.159	–2741
		(210.5)		(.040)	(1324)
1970	2758	–233.8			–1470
		(39.7)			(250)
1969	2299	–2.0			
		(34.5)			

Notes: Figures are in nominal US dollars. There are about 13,500 observations with earnings in each cohort. Standard errors are shown in parentheses

Sources: Adapted from Angrist (1990, Tables 2 and 3), and unpublished author tabulations. Earnings data are from Social Security administrative records. Veteran status data are from the Survey of Income and Program Participation

average Social Security-taxable earnings in column (2). Column (1) shows average annual earnings for purposes of comparison. For men born in 1950, there are significant negative effects of eligibility status on earnings in 1970, when these men were being drafted, and in 1981, ten years later. For example, the 1981 estimate for whites is – 436 dollars. In contrast, there is no evidence of an association between eligibility status and earnings in 1969, the year the lottery drawing for men born in 1950 was held but before anyone born in 1950 was actually drafted.

Because eligibility status was randomly assigned, the claim that the estimates in column (2) represent the causal effect of *draft eligibility* on earnings seems uncontroversial. An additional assumption embodied in Eq. (6) is that the only reason eligibility affects earnings is military service. Given this, the only information required to go from draft-eligibility effects to veteran-status effects is the denominator of the Wald estimator, which is the effect of draft-eligibility on the probability of serving in the military. This information is reported in column (4) of Table 2, which shows that draft-eligible men were 0.16 more likely to have served in the Vietnam era. For earnings in 1981, long after most Vietnam-era servicemen were discharged from the military, the Wald estimates of the effect of military service reported in column (5) amount to about 15 percent of earnings. Effects were even larger in 1970, when affected soldiers were still in the army.

IV with Heterogeneous Treatment Effects

The constant-effects assumption is clearly unrealistic. We'd like to allow for the fact that some men may have benefited from military service while others were undoubtedly hurt by it. In general, however, IV methods fail to capture either ATE or ATET in a model with heterogeneous treatment effects. Intuitively, this is because only a subset of the population is affected by any particular instrumental variable. In the draft lottery example, many men with high lottery numbers volunteered for service anyway (indeed, most Vietnam veterans were volunteers), while many draft-eligible men nevertheless avoided service. The draft lottery instrument is not informative about the effects of military service on men who were unaffected by their draft-eligibility status. On the other hand, there is a sub-population who served solely because they were draft-eligible, but would not have served otherwise. Angrist, Imbens and Rubin (1996) call the population of men whose treatment status can be manipulated by an instrumental variable the set of *compliers.* This term comes from an analogy to a medical trial with imperfect compliance. The set of compliers are those who 'take their medicine', that is, they serve in the military when draft-eligible but they do not serve otherwise.

Under reasonably general assumptions, IV methods can be relied on to capture the causal effect of treatment on compliers. The average causal effect for this group is called a local average treatment effect (LATE), and was first discussed by Imbens and Angrist (1994). A formal description of LATE requires one more bit of notation. Define potential treatment assignments D_{0i} and D_{1i} to be individual i's treatment status when Z_i equals 0 or 1. One of D_{0i} or D_{1i} is counterfactual since observed treatment status is

$$D_i = D_{0i} + Z_i(D_{1i} - D_{0i}).$$

The key identifying assumptions in this setup are (a) conditional independence, that is, that the joint distribution of $\{Y_{1i}, Y_{0i}, D_{1i}, D_{0i}\}$ is independent of Z_i; and (b) monotonicity, which requires that either $D_{1i}D_{0i}$ for all i or vice versa. Monotonicity requires that, while the instrument might have no effect on some individuals, all of those who are affected should be affected in the same way (for example, draft eligibility can only make military service more likely, not less). Assume without loss of generality that monotonicity holds with $D_{1i} \geq D_{0i}$. Given these two assumptions, the Wald estimator consistently estimates LATE, written formally as $E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}]$. In the draft lottery example, this is the effect of military service on those veterans who served because they were draft eligible but would not have served otherwise. In general, LATE compliers are a subset of the treated. An important

special case where LATE = ATET is when D_{0i} equals zero for everyone. This happens in a social experiment with imperfect compliance in the treated group and no one treated in the control group.

IV Details

Typically, covariates play a role in IV models, either because the IV identification assumptions are more plausible conditional on covariates or because of statistical efficiency gains. Linear IV models with covariates can be estimated most easily by two-stage least squares (2SLS), which can also be used to estimate models with multi-valued, continuous, or multiple instruments. See Angrist and Imbens (1995) or Angrist and Krueger (2001) for details and additional references.

See Also

- ▶ Instrumental Variables
- ▶ Matching Estimators
- ▶ Regression-Discontinuity Analysis
- ▶ Rubin Causal Model
- ▶ Selection Bias and Self-Selection
- ▶ Two-Stage Least Squares and the k-Class Estimator

Bibliography

Angrist, J. 1990. Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review* 80: 313–335.

Angrist, J. 1991. Grouped-data estimation and testing in simple labor-supply models. *Journal of Econometrics* 47: 243–266.

Angrist, J. 1998. Estimating the labor market impact of voluntary military service using Social Security data on military applicants. *Econometrica* 66: 249–288.

Angrist, J., and G. Imbens. 1995. Two-stage least squares estimates of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 90: 431–442.

Angrist, J., G. Imbens, and D. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91: 444–455.

Angrist, J., and A. Krueger. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives* 15(4): 69–85.

Ashenfelter, O. 1978. Estimating the effect of training programs on earnings. *Review of Economics and Statistics* 6: 47–57.

Ashenfelter, O., and D. Card. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics* 67: 648–660.

Dehejia, R., and S. Wahba. 1999. Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94: 1053–1062.

Heckman, J. 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475–492.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.

Heckman, James J., and R. Robb. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, ed. J. Heckman and B. Singer. New York: Cambridge University Press.

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81: 945–970.

Imbens, G., and J. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62: 467–475.

Quandt, R. 1958. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* 53: 873–880.

Rosenbaum, P., and D. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66: 688–701.

Rubin, D. 1977. Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* 2: 1–26.

Wald, A. 1940. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* 11: 284–300.

T

Trend/Cycle Decomposition

Charles R. Nelson

Abstract

The decomposition of economic time series is motivated by the idea that distinct forces

account for long-term growth, for variation over a time frame associated with the business cycle, and though the seasons. While the latter is typically suppressed by 'seasonal adjustment', the issue of how to separate trend from cycle in series such as GDP has been hotly debated since the 1970s and remains unsettled. Surprisingly varied patterns follow from alternative approaches, some placing the bulk of variation into the cycle/trend following a smooth line, others attributing shifts in level to 'permanent shocks' to the trend.

Keywords

ARMA models; Business cycles; Ergodicity; Filtering; Identification; Kalman filter; Random walks; Trend/cycle decomposition; Unobserved components models

JEL Classifications

C1

Macroeconomists distinguish between the forces that cause long-term growth and those that cause temporary fluctuations such as recessions. The former include population growth, capital accumulation, and productivity change, and their effect on the economy is permanent. The latter are generally monetary shocks such as shifts in central bank policy that affect the real economy through price rigidities that cause output to deviate temporarily from its long-run path. This conceptual dichotomy motivates the decomposition of aggregate output, real GDP, into two components: the *trend* which accounts for long-term change, and the *cycle* which is a short-term deviation from trend. While economists no longer believe the 'business cycle' to be deterministically periodic, that terminology remains. Seasonal variation could be a third component, though it has been suppressed in 'seasonally adjusted' data such as GDP.

This suggests we may express the natural log of GDP (or any other 'trending' time series), denoting the observation at time t by 'y_t', as follows:

$$y_t = \tau_t + c_t.$$

Here τ_t denotes the value of the trend and c_t the cycle at time t, neither of which is observed directly. Since this single equation cannot be solved directly for the unknown trend and cycle, additional assumptions are required for 'identification', a procedure which allows estimates of them to be calculated from the GDP data. The fundamental identifying assumption is that the cycle component is temporary, that it dies out after a sufficiently long time. However, this assumption of 'stationarity' or 'ergodicity', which distinguishes it from trend, which is permanent, does not suffice by itself to achieve identification. More has to be said about the nature of the trend.

The simplest specification of trend is to make τ_t a linear function of time where the slope is the long-term growth rate. A second identifying assumption is that trend should account for as much of the variation in the data as possible, minimizing the amplitude of the implied cycle. This is achieved by least squares regression of y_t on time and the estimated trend is $\widehat{\tau}_t = a + b\ time$ where a and b are estimates of intercept and slope respectively. The implied cycle component is then $\widehat{c}_t \equiv y_t + \widehat{\tau}_t$. Though successful in accounting for a large fraction of the change in GDP over long periods, this approach implies cycles of extraordinary length, well beyond the roughly seven years between recession dates identified by the National Bureau of Economic Research for the United States, and the pattern is contrary to economic intuition (for the United States the 1970s, a decade of poor economic performance, were well above the trend line while the 1990s, a decade of prosperity, were well below trend). A more flexible trend function is clearly called for, but quadratic or higher-order polynomials in time imply unstable paths when extrapolated into the future. Perron (1989) suggested a segmented trend function allowing for an occasional change level or slope to be captured by dummy variables.

A general approach to estimating a flexible and adaptive trend is filtering, where estimated trend is a weighted average of adjacent observations. Here it is the weighting scheme which identifies the components. For example, $\widehat{\tau}_t = .25 \cdot y_{t-1} + .50 \cdot y_t + .25 \cdot y_{t+1}$ applies symmetric though

unequal weights to the current observations and its immediate neighbours. No filter is perfect in the sense of revealing the actual trend, but a desirable filter is one that extracts as much of the trend as possible from the data. A criterion for choosing a filter would be that it produces cycles having characteristics that match our notions of the business cycle, for example that recessions occur on average about every seven years. A widely used filter that does this is the Hodrick and Prescott (1980), filter which penalizes deviations from trend and changes in trend through a loss function.

The distinction between trend and cycle implies that the forecast of GDP far in the future must be the trend, since the cycle will die away. The approach to trend/cycle decomposition proposed by Beveridge and Nelson (1981) turns this conclusion on its head by proposing that the trend at a date in time be defined as the forecast of the distant future (adjusted for average growth). Specifically, they estimate an autoregressive moving average (ARMA) time series model for the growth rate and compute the forecast of the level into the distant future, adjusting for average growth. The resulting measure of trend shows whether actual GDP is above or below its forecast growth path, the difference being the cycle. Since parameters of the ARMA model are identified, and computation of forecasts is straightforward, the Beveridge–Nelson decomposition is identified. It turns out that the trend component is a random walk with drift regardless of the specific ARMA model, and this accords with the intuition that only unexpected shocks can affect a long horizon forecast.

To obtain the general expressions for the components we rearrange the ARMA model as:

$$\varphi(L)\Delta y_t = \theta(L)\varepsilon_t$$
$$\Delta y_t = \psi(L)\varepsilon_t$$

where the average growth rate has been suppressed, the statistical shock ε_t is serially random, Δ denoted first difference, and L is the lag operator, and $\psi(L) = \theta(L)/\varphi(L)$. The growth rate of GDP can be thought of as a weighted history of all past shocks where the coefficient of ε_{t-k} is ψ_k plus the expected average growth rate μ. It may be shown that an algebraically equivalent expression is

$$y_t = \psi(1)\sum_{k=0}^{\infty}\varepsilon_{t-k} - \widetilde{\psi}(L)\varepsilon_t \quad \widetilde{\psi}_k = \sum_{j=k+1}^{\infty}\widetilde{\psi}_j.$$

Note that the first term is the sum of all past shocks each with weight equal to the total effect of all past shocks. The second term may be shown to be a stationary time series with mean zero. Thus the trend is always a random walk regardless of the ARMA model.

For example, growth in US GDP is roughly an AR(1) process with coefficient .25, so the effect of a shock on the trend is $\psi(1) = 1/(1 - .25) = 1.33$. This illustrates the surprising implication that the trend component may be highly variable; indeed, the results obtained by Beveridge and Nelson imply that variation in observed GDP is largely the result of variation in the trend component and is therefore permanent.

Unobserved components models identify trend and cycle by specifying a separate and specific stochastic process for each. The trend is generally assumed to be a random walk with drift, allowing it to account for long-term growth while permitting it to be shifted by stochastic shocks. The cycle is assumed to be a process that is stationary in the sense of reverting to a mean over time. (The mean of the cycle is zero for symmetric variation around trend, but evidence exits for asymmetric cycles with a negative mean.) This approach was introduced to economics by Harvey (1985) and Clark (1987). An example would be the following:

$$\tau_t = \tau_{t-1} + \mu + \eta_t \quad c_t = \varphi \cdot c_{t-1} + \varepsilon_t.$$

The parameter μ is the long-term growth rate, the shock η is random and may be positive or negative, the parameter φ measures the persistence of the cycle, and shocks ε drive the cycle. The two shocks are often assumed to be uncorrelated, which reduces the number of parameters to be estimated by one but may also place an unwarranted restriction on the relation between the two components. More generally the cycle process may have a higher-order ARMA specification. Identification of the parameters depends on whether a specific model implies a sufficient number of estimable parameters in the corresponding ARMA reduced form representation of

Δy_t (corresponding to identification of simultaneous equation models). Given an identified model and parameter estimates, the estimated trend and cycle may be computed using the Kalman filter.

A useful result is that the random walk trend in the unobserved components model is identified even if its parameters are not identified. Morley et al. (2003) show that the Beveridge–Nelson trend is always the conditional expectation of the trend component given past data. What identifies the trend is the random walk specification for the trend along with the assumption that the cycle process does not persist indefinitely. Thus, the long-horizon forecast reflects only the trend, and such forecasts can always be computed from the reduced form ARMA model.

See Also

- ▶ Business Cycle Measurement
- ▶ Data Filters
- ▶ State Space Models
- ▶ Time Series Analysis
- ▶ Unit Roots

Bibliography

Beveridge, S., and C.R. Nelson. 1981. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. *Journal of Monetary Economics* 7: 151–174.

Clark, P.K. 1987. The cyclical component of U.S. economic activity. *Quarterly Journal of Economics* 102: 797–814.

Harvey, A.C. 1985. Trends and cycles in macroeconomic time series. *Journal of Business and Economic Statistics* 3: 216–227.

Hodrick, R.J., and E.C. Prescott. 1980. *Postwar US business cycles: An empirical investigation*, Working paper. Pittsburgh: Carnegie-Mellon University.

Hodrick, R.J., and E.C. Prescott. 1997. Postwar U.S. business cycles: An empirical investigation. *Journal of Money Credit and Banking* 29: 1–16.

Morley, J.C., C.R. Nelson, and E. Zivot. 2003. Why are the Beveridge–Nelson and unobserved-components decompositions of GDP so different? *Review of Economics and Statistics* 85: 235–243.

Perron, P. 1989. The great crash, the oil price shock, and the unit root hypothesis. *Econometrica* 57: 1361–1401.

Triffin, Robert (1911–1993)

Albert Kervyn

Keywords

European Monetary System; European Payment Union; Gold-exchange standard; International Monetary Fund; International reserves; Special Drawing Rights; Triffin Plan; Triffin, R

JEL Classifications

B31

Triffin was born in 1911 in the pleasant Belgian village of Flobecq. After a brilliant school career, a scholarship enabled him to study law and economics at the university of Louvain. Another scholarship sent him to Harvard, where he got a thorough grounding in theory with Schumpeter and Leontief. His 1938 dissertation, 'Monopolistic Competition and General Equilibrium Theory', earned a Wells Prize and was published in 1940.

After a brief return to Belgium, he was appointed instructor at Harvard, and was soon cut off from Europe by the outbreak of the war. In 1942 he joined the Federal Reserve Board to organize a research section on Latin America. This launched him on his parallel career of advising central banks on reform of monetary and exchange arrangements. He rapidly developed what was to remain a main characteristic of his work in this area: a flair for practical suggestions and an imagination that provided alternatives in case of political objections to the first proposals. His success in a number of Latin American countries led to his appointment in 1946 as head of the exchange control division in the newly-created International Monetary Fund (IMF). Moving to Europe as IMF chief representative, he developed a proposal for the European Clearing Union. Having later transferred to the State Department, he succeeded in negotiating his

proposal through OEEC, and became the recognized father of the European Payment Union of 1950.

Triffin left the State Department after a policy disagreement, and went to Yale, where he stayed from 1951 to 1977. There he published his two classic works: *Europe and the Money Muddle* in 1957; *Gold and the Dollar Crisis* in 1959 and 1960. The first book reviews the European monetary experience, using an integrated analysis of the money supply and the balance of payments. It proclaims the end of the 'dollar shortage' and explains the dilemma of the gold-exchange standard in the absence of an adequate supply of gold: either the key-currency country maintains equilibrium in its balance of payments, and other countries will experience a shortage of the reserves needed to support an expansion of trade and transactions, with an attendant brake on growth; or the desirable growth of world reserves will be preserved only through persistent increases in the liabilities of the key-currency country, raising increasing doubts about its ability to redeem such liabilities, especially when they begin to exceed the country's dwindling gold reserves. The second book, shorter and less analytic, goes further: it contains a bold prophecy of a dollar glut, bound in time to bring down the gold-exchange standard. The fate experienced by sterling in the interwar years was bound to catch up with the dollar as well.

The second part of the book contained the famous 'Triffin Plan' to obviate these dangers: on the one hand, the controlled creation of an international reserve instrument by the IMF; on the other hand, regional monetary arrangements, with emphasis on European integration.

The announced dollar glut did indeed develop over the following years with the predicted consequences; the gold parity rate was first abandoned, then the dollar became inconvertible. The first policy prescription was timidly followed with the creation of Special Drawing Rights, but without the essential element of conversion of dollar balances. Indeed the excessive accumulation of such balances was later denounced by Triffin as the main factor in the inflationary development of the 1970s, and as an unwarranted capital inflow into the USA, which on the contrary should be a leading exporter of capital.

The other prescription was more successful, and Triffin moved back to Europe (and the University of Louvain) in order to participate more fully in the emergence and development of the European Monetary System.

He was throughout a most active participant in the debate on money and exchanges. In a continuous stream of studies, papers and memoranda, he pressed forward for advances on the two fronts of international monetary reform and European integration.

His reputation as an analyst and skilled deviser of techniques led to his being called in as a consultant all over the world, on domestic as well as regional monetary issues.

Later in life, Triffin remained a much concerned 'citizen of the world' and applied his policy-oriented brand of 'economics of persuasion' not only to his particular area of expertise, but also to the issues of development and disarmament.

Selected Works

As a policy-oriented persuader, Triffin has been a prolific writer. In his attempt to reach a wide public for his ideas, he gives the impression of never having refused a contribution. As a result a complete bibliography would include well over 300 items – many in less accessible publications. The following list is therefore highly selective.

1940. *Monopolistic competition and general equilibrium theory.* Cambridge, MA: Harvard University Press.

1957. *Europe and the money Muddle.* New Haven: Yale University Press.

1960a. *Statistics of sources and uses of finance, 1948–58.* Paris: OEEC.

1960b. *Gold and the dollar crisis.* New Haven: Yale University Press.

1964. *The evolution of the international monetary system.* Princeton studies in international

finance no. 12. Princeton: Princeton University Press.
1966. *The world money maze.* New Haven: Yale University Press.
1969. The thrust of history in international monetary reform. *Foreign Affairs* 47: 477–492.
1971. The use of SDR finance for collectively agreed purposes. *Banca Nazionale del Lavoro Quarterly Review* 96(March): 3–12.
1973. The role of developing European monetary union. In *Europe and the evolution of the international monetary system,* ed. A.K. Swoboda. Leiden: Sijthoff.
1975. The Community and the disruption of the world monetary system. *Banca Nazionale del Lavoro Quarterly Review* 112(March): 3–35.
1976. Size, sources and beneficiaries of international reserve creation, 1970–74. In *Economic progress, private values and public policy*, ed. B. Balassa and R. Nelson. New Haven: Yale University Press.
1978a. 'Europe and the money muddle' revisited. *Banca Nazionale del Lavoro Quarterly Review* 124(March): 49–65.
1978b. *Gold and the dollar crisis*: *Yesterday and today.* Princeton essays in international finance no. 132. Princeton: Princeton University Press.
1978c. The international role and fate of the dollar. *Foreign Affairs* 57: 269–286.
1981a. Le système monétaire Européen dans le cadre du système monétaire mondial. *Banque* 55, 535–40.
1981b. An economist's career: What? Why? How? *Banca Nazionale del Lavoro Quarterly Review* 139(September): 239–259.
1981c. The first two years of FECOM transactions. Economic papers no. 2. Brussels: Directorate-General for Economic and Financial Affairs, European Commission, July.
1984a. How to end the world 'infession'. In *Europe's money,* ed. R.S. Masera and R. Triffin. Oxford: Clarendon Press.
1984b. Une tardive autopsie du Plan Keynes de 1943. *Revue d'économie politique.*
1986. Correcting the world monetary scandal. *Challenge* 28(January/February): 4–14.

Trotsky, Lev Davidovitch (1879–1940)

Richard B. Day

Keywords

Bukharin, N. I.; Collectivization; Comparative coefficients; Industrialization; Planning; Socialism; Stalin, J. V.; Trade monopoly; Trade unions; Trotsky, L. D

JEL Classifications

B31

Born in 1879, the son of Jewish farmers living near the Black Sea, Trotsky became an important political figure by the time of the Second Congress of the Russian Social Democratic Party in 1903. Disagreeing with Lenin's centralizing view of party organization, Trotsky either favoured the Mensheviks or attempted to mediate between them and the Bolsheviks until making his peace with Lenin in 1917. In the 1905 Revolution he served as chairman of the St Petersburg Soviet, drawing upon that experience to develop the theory of 'permanent revolution' in his book *Results and Prospects.* In the 1917 Revolution Trotsky ranked second only to Lenin among Bolshevik party leaders. He orchestrated the seizure of power and subsequently organized and led the Red Army in the civil war. During the early 1920s Trotsky's political influence waned, and by the middle of the decade he became the political leader and intellectual mentor of the Left Opposition to Stalin. Defeated by Stalin in the intra-party struggle, in 1929 Trotsky was deported from the Soviet Union. In exile he edited *Biulleten' Oppozitsii* (Bulletin of the Opposition) and published numerous other writings critical of Stalinist policy, the most important being *The Revolution Betrayed.* Unable to answer Trotsky's criticisms on intellectual grounds, in August 1940 Stalin replied in the only way he knew: he had Trotsky assassinated in Mexico, his last place of exile.

In *Results and Prospects* (first published in 1906), Trotsky predicted that Russian backwardness would guarantee the revolution in permanence. Surrounded by stronger enemies, the Russian state had prevented the nobility from becoming politically independent. The nobility were mere tax collectors, extracting revenue from the peasants in order to promote development; and the bourgeoisie, likewise, were weaker than their Western counterparts, for much of the economy was built with foreign loans, serviced by grain exports. The proletariat, in contrast, enjoyed disproportionate strength. Few in number, Russian workers were concentrated in large factories organized around foreign technology. Trotsky predicted that the proletariat would overthrow the autocracy, by-passing the bourgeois revolution, but would then confront a counter-revolutionary alliance when it implemented its programme. The counter-revolution would be supported by Germany, Austria and France, who would be anxious to prevent the revolution's spread and to safeguard their investments. When these countries mobilized, however, they would drive their own workers to revolt, thereby making the revolution permanent both domestically and internationally.

Aware of Russia's historical dependence on the world economy, Trotsky characteristically viewed economic issues in an international context. Modern industry, he believed, had become so capital intensive that production could only be profitable through specialization in service of the world market. It was in the nature of socialism to emancipate the productive forces from the fetters of the nation state. A victory of the proletariat in the leading countries would mean 'a radical restructuring of the very economic foundation in correspondence with a more productive international division of labour, which is alone capable of creating a genuine foundation for a socialist order' (Trotsky Archives No. T-3148).

When the international revolution did not come to Soviet Russia's aid as Trotsky had expected, he continued to insist that industrialization must draw upon the resources of the world market. Opposing Stalin's notion of an isolated socialist state (Socialism in One Country), he argued that 'a properly regulated growth of export and import with the capitalist countries prepares the elements of the future commodity and product exchange [which will prevail] when the European proletariat assumes power and controls production' (Trotsky Archives No. T-3034). Soviet Russia's relation to the West would involve a dialectic of cooperation and struggle in which the Soviet state would regulate its 'dependence' on capitalism through its monopoly of foreign trade. The alternative, the Stalinist vision of autarky, would mean reliance 'on the curbed and domesticated productive forces, that is ... on the technology of backwardness' (Trotsky 1941, p. 53).

Uppermost in Trotsky's mind throughout the 1920s was the need not only to preserve access to foreign technology, but also to reduce domestic prices in order to maintain the trade monopoly. In 1923 he warned the party that 'Contraband is inevitable if the difference between external and internal prices goes beyond a certain limit ... contraband, comrades ... undermines and washes away the monopoly' (*Dvenadstatyi s 'ezd RKP* (b) (1923), p. 372; 12th Congress of the Russian Communist Party (Bolsheviks)). Without this protection for new Soviet industries, planned growth would be impossible.

For the promotion of new industrial construction, Trotsky proposed to supplement domestic tax revenues by taking advantage of Europe's need for foreign markets and by pursuing all manner of credits:

> What does foreign credit do for our economic development? Capitalism makes advances to us against our savings which do not yet exist ... Asa result, the foundations of our development are extended ... The dialectics of historical development have resulted in capitalism becoming for a time the creditor of socialism. Well, has not capitalism been nourished at the breasts of feudalism? History has honoured the debt. (*Pravda,* 20 September 1925)

In addition to making use of foreign credits, Trotsky hoped to resume the tsarist pattern of exporting grain in exchange for finished goods. In 1925 he predicted that the Soviet economy would be unable to satisfy more than a fraction of its need for new equipment:

> We must not... forget for a moment the great mutual dependence which existed between the economies of tsarist Russia and world capital. We must just bring to mind the fact that nearly two-thirds of the technical equipment in our works and factories used to be imported from abroad. This dependence has hardly decreased in our own time, which means that it will scarcely be economically profitable for us in the next few years to produce at home the machinery we require, at any rate, more than two-fifths of the quantity, or at best more than half of it. (*Pravda,* 20 September 1925)

Trotsky hoped to reconcile a high level of foreign trade with socialist protectionism through strict determination of priorities. Soviet industries should economize on scarce capital, specialize in those products in greatest demand, standardize output and reduce costs, while leaving the remaining needs to be met by low-cost imports. A system of comparative coefficients should be devised by the planners, comparing the cost and quality of Soviet products with foreign competition. A poor coefficient would then signal the advisability of imports in the short run and of re-equipment in the long run, as new resources became available. 'A comparative coefficient is the same for us as a pressure gauge for a mechanic on a locomotive. The pressure of foreign production is for us the basic factor of our economic existence. If our relation to this production is [unsatisfactory], then foreign production will sooner or later pierce the trade monopoly' (*Ekonomicheskaia Zhizn',* 18 August 1925).

In spite of his balanced approach to industrialization, official Soviet historiography insists that Trotsky was a 'super-industrializer', determined to plunder the peasantry. In reality he attempted more systematically than any of his contemporaries to avert the crisis of forced industrialization by balancing the needs of the peasantry against those of industry through a policy of 'commodity intervention'. To the extent that export-oriented growth clearly depended upon the peasants bringing grain to market, Trotsky was quite aware that the most urgent consumer needs would also have to be satisfied through imports. The world market was to function as a 'reserve' for both light and heavy industry. The 'goods famine', or the chronic shortage of consumer goods, was 'obvious and incontestable proof that the distribution of national economic resources between state industry and the rest of the economy has ... acquired the necessary proportionality' (Trotsky Archives No. T-2983). The real enemies of the peasantry, in Trotsky's view, were the authors of Socialism in One Country – Stalin, who saw only the needs of the machine-building industries, and Bukharin, who urged the peasant to 'enrich' himself without seriously considering the need to provide consumer goods upon which these savings might be spent.

It was Trotsky's concern for the legitimate needs of workers and peasants alike which led him in the 1930s to reconsider the role of market forces, for a time at least, in socialist planning. As early as 1925 he had warned that it was 'impossible to push industrialization forward with the aid of unreal credits' (Trotsky 1955, p. 186). During the first five-year plan he called for restraints upon the inflationary financing of heavy industry and 'strict financial discipline', even at the expense of closing down enterprises. A stable currency, in turn, would provide an instrument whereby the masses themselves could democratically control production decisions from below. 'The innumerable living participants in the economy', Trotsky wrote in 1932,

> state and private, collective and individual, must announce their needs and their respective intensities not only through the statistical calculations of the planning commissions, but also by the direct pressure of supply and demand. The plan ... [must be] verified, and in an important measure must be achieved through the market. (*Biulleten' Oppozitsii* 31 (1932), p. 8)

A planned market, free trade unions, and restoration of Soviet democracy: these were the three elements without which any talk of socialism was a mockery.

> If there existed the universal mind described in the scientific fantasy of Laplace – a mind which might simultaneously register all the processes of nature and society, measure the dynamic of their movement and forecast the results of their interactions – then, of course, such a mind could *a priori* draw up a faultless and exhaustive economic plan, beginning with the number of hectares of wheat and ending with buttons on a waistcoat. True, it often appears to the bureaucracy that it possesses just such a mind:

> and that is why it so easily emancipates itself from control by the market and by soviet democracy. The reality is that the bureaucracy is cruelly mistaken in its appraisal of its own spiritual resources. (*Biulleten' Oppozitsii* 31 (1932), p. 8)

In *The Revolution Betrayed* (1937), his most thorough critique of Stalinist 'planomania', Trotsky concluded that the real basis of bureaucratic power had nothing to do with Stalin's pompous claims of industrial triumphs; the horrible truth was that the whole bureaucratic edifice had come to rest upon nothing more profound or despicable than an ability to manufacture poverty. Queues were the foundation of Soviet power and the innermost secret of the police state:

> The basis of bureaucratic rule is the poverty of society in objects of consumption. When there are enough goods in a store, the purchasers can come whenever they want to. When there are few goods, the purchasers are compelled to stand in line. When the lines are very long, it is necessary to appoint a policeman to keep order. Such is the starting point of the Soviet bureaucracy. It 'knows' who is to get something and who has to wait. (Trotsky 1937, p. 112.

Historians will continue to debate whether Trotsky's policies might have avoided forced collectivization and the excesses of Stalin's five-year plans. On one point, however, there can be no dispute: Trotsky was perfectly correct to conclude that Stalin's pursuit of autarky had more in common with the ideals of Hitler than with those of Marx. The Russian revolution, confined to a single backward country, did not lead to the emancipation of the proletariat. Trotsky attempted to reinterpret and apply Marxism to the unexpected conditions of an isolated revolutionary experiment. He did not win the battle against Stalin. He did, however, help to explain and attempt to avert one of the great tragedies of the twentieth century.

Selected Works

1921. *Terrorism and communism.* Ann Arbor: University of Michigan Press, 1963.
1923. *The new course.* Ann Arbor: University of Michigan Press, 1965.
1926. *Towards socialism or capitalism?* London: Methuen & Co.
1927. *The platform of the left opposition.* London: New Park Publications, 1963.
1930a. *My life.* New York: Grosset & Dunlap, 1960.
1930b. *The history of the Russian revolution.* Trans. M. Eastman. London: Victor Gollancz, 1965.
1937. *The revolution betrayed.* New York: Pioneer Publishers, 1945.
1941. *Stalin: An appraisal of the man and his influence*. London: Hollis & Carter, 1947.
1962. *The permanent revolution and results and prospects.* Trans. J.G. Wright and B. Pearce. London: New Park Publications.
1973. *Biulleten' Oppozitsii* (Bulletin of the Opposition). New York: Monad.

Bibliography

Day, R.B. 1973. *Leon Trotsky and the politics of economic isolation*. London: Cambridge University Press.
Deutscher, I. 1954. *The prophet armed. Trotsky: 1879–1921*. London: Oxford University Press.
Deutscher, I. 1959. *The prophet unarmed. Trotsky: 1921–1929*. London: Oxford University Press.
Deutscher, I. 1963. *The prophet outcast. Trotsky: 1929–1940*. London: Oxford University Press.
Figes, O. 1996. *A people's tragedy: A history of the Russian revolution*. New York: Viking Press.
Howe, I. 1978. *Leon Trotsky.* New York: Viking.
Knez-Paiz, B. 1978. *The social and political thought of Leon Trotsky.* Oxford: Clarendon Press.
Thatcher, I. 2003. *Trotsky.* London: Routledge.

Troubled Asset Relief Program (TARP)

Linus Wilson

The Troubled Asset Relief Program (TARP), or the $700 billion bailout, has been the subject of much academic interest. Here the rigorous studies on the programs of this massive intervention into the financial sector are reviewed. While

considerable work has been done on the bank bailouts in the TARP, the troubled asset programs, automotive rescues, homeowner assistance programs, and ad hoc bailouts have not been subjected to much theoretical modelling or empirical research.

Background of the TARP

In 2008, the Federal Reserve stretched its powers to lend to businesses and individuals in unusual and exigent circumstances. It extended over $100 billion in loans to assist the purchase of the investment bank Bear Sterns by JPMorgan Chase and to prop up American International Group (AIG). Yet the Federal Reserve lacked the powers to inject capital into troubled but systemically important institutions. AIG, which guaranteed the mortgage-backed securities of many of the global investment banks, was without doubt one such institution. Its failure could have caused a cascade of failures of global investment houses, such as Goldman Sachs, UBS and Credit Suisse. After the bankruptcy filing of Lehman Brothers and the seizure of mortgage giants Fannie Mae and Freddie Mac by the federal government in September 2008, funding markets froze and the failures of the large, highly respected, independent investment banks Morgan Stanley and Goldman Sachs seemed possible.

On 15 April 2008, in the US Treasury department, two low-level officials developed a 'break the glass' memo to buy hundreds of billions of dollars' worth of troubled real estate-backed assets ('toxic' assets) from banks. This plan was proposed to Congress by former Goldman Sachs CEO and then US Treasury Secretary Henry 'Hank'; Paulson with the support of President George W. Bush and Federal Reserve Chairman Ben Bernanke. On 20 September 2008, they asked for $700 billion in a three-page bill in part because $700 billion was the most they could dream of getting from Congress. The House voted the bill down on 29 September 2008. Then, on 1 October 2008, the Senate passed the bill. On 3 October 2008, the House reconsidered and passed a much larger Emergency Economic Stabilization Act (ESSA) to authorise the $700 billion Troubled Asset Relief Program (TARP). (There are several studies on the factors associated with the various votes on the EESA legislation. See Green and Hudak (2009), Ramı´ rez (2011), Wen (2011) and Wurtz (2010). Ramı´ rez (2011) finds that vote switching in the House of Representatives from against to in favour of the TARP legislation was positively related to political action committee (PAC) contributions from the American Bankers Association (ABA), a trade group of commercial banks. The timeline is taken from Propublica.org (2013).) On the same day, President Bush signed the bill that gave the Secretary of the US Treasury authority to purchase all kinds of financial assets. The TARP authority to allocate money to asset purchases expired on 3 October 2010, but already allocated funds are still being spent. No more than $467.2 billion will be expended by the program. Most of those expenditures have been returned. Table 1's repayment figures do not include the dividends, interest and capital gains from the TARP investments.

While even the best investment returns of the TARP have lagged the risk-adjusted cost of capital at the time of purchase, taxpayers stand a good chance of earning a nominal profit on the TARP as a whole. Moreover, the TARP, special FDIC deposit and debt guarantees (For a description of the FDIC debt and emergency deposit guarantee programs, see Wilson and Wu (2011).), and trillions of dollars of Federal Reserve emergency lending programs were successful in stabilising large systemically important institutions during a potentially devastating financial crisis. Yet these interventions still have the potential of encouraging reckless behaviour by large financial institutions in the future.

Here we look at the academic work in economics, finance and related business disciplines into the major programs of the TARP. The emphasis of this review is on studies that empirically tested the TARP or used explicit mathematical modelling of TARP's programs. The many good policy discussion papers which were important in framing the debate about the proper structure and role of TARP are generally excluded from our discussion. With the

Troubled Asset Relief Program (TARP), Table 1 The programs of the TARP and expenditures in billions of US dollars through 30 June 2012

TARP Program	Type of assistance	Current obligation	Expenditure	Principal repaid	Still owed to taxpayers	Available to be spent
Capital Purchase Program	Preferred stock, common stock, and subordinated debt investments in banks and credit unions	204.9	204.9	191.1	13.8	0.0
Automotive Industry Support Programs	Loans, guarantees, and capital for automakers, auto finance companies, and suppliers	79.7	79.7	35.2	44.5	0.0
Systemically Significant Failing Institutions	Preferred and common stock investments in AIG	67.8	67.8	31.9	36.0	0.0
Housing Support Programs	Payments to servicers to modify loans of struggling homeowners	45.6	4.5	0.0	0.0	41.1
Targeted Investment Program	Preferred stock investments in Bank of America and Citigroup	40.0	40.0	40.0	0.0	0.0
Public Private Investment Program	Non-recourse loans to and equity stakes in investment funds purchasing residential and commercial mortgage backed securities	21.9	18.5	4.4	14.1	3.4
Asset Guarantee Program	Asset guarantees for Bank of America and Citigroup	5.0	0.0	0.0	0.0	0.0
Term Asset-Backed Loan Facility	Subordinated debt investments to finance the purchase of securitised ssets	1.4	0.1	0.0	0.1	1.3
Community Development Capital Initiative	Preferred stock and subordinated debt investments in small banks and credit unions	0.6	0.2	0.0	0.6	0.0
Unlocking Credit for Small Businesses	Purchases of securitised small business loans	0.4	0.4	0.4	0.0	0.0
	Totals	**467.3**	**416.1**	**302.9**	**109.1**	**45.8**

Source: SIGTARP (2012, p. 40) and author's analysis

exception of the Capital Purchase Program (CPP) for 'healthy' banks, most of the programs of the TARP have been largely ignored in terms of rigorous research. Hopefully, future scholarship will correct this deficit. We will discuss the bank bailouts, troubled asset purchase programs, the auto bailouts, the homeowner assistance programs and the large ad hoc bailouts of AIG, Bank of America and Citigroup in turn.

The Well-Studied Bank Bailout: The Capital Purchase Program

Despite having troubled assets in its name, the TARP's first expenditures were to buy bank stock, not troubled assets. Top executives of nine major banks were cajoled in a room of top regulators to accept $125 billion in exchange for preferred stock and warrants on the Columbus Day

holiday of 2008 (Those banks with the first TARP capital infusions were Bank of America, Bank of New York Mellon, Citigroup, Goldman Sachs, JPMorgan Chase, Morgan Stanley, Merrill Lynch, State Street and Wells Fargo.). Before the Capital Purchase Program (CPP) closed at the end of 2009, 707 banks accepted about $205 billion from taxpayers.

Most of the money was given in exchange for preferred stock that paid 5% per annum for the first 5 years and 9% per annum thereafter. Publicly held banks also issued common stock warrants to taxpayers, according to Wilson (2009b). Privately held banks issued immediately exercised preferred stock warrants with a par value of 5% of the investment amount. Those private bank preferred stock warrants paid a 9% dividend per annum from the date of exercise. Subordinated debt was issued by S-corp banks at a rate of 7.8% for the first 5 years and 13.8% thereafter.

It is not clear from the theoretical literature whether the preferred stock investments could have encouraged banks to increase their underwriting of good loans. Wilson (2009a) and Wilson and Wu (2010) argue that preferred shares from the CPP would have done little to address the problems associated with excessive leverage. Namely, banks with too much leverage and too little common equity capital will pass up safe but profitable lending opportunities due to debt overhang, and will seek out unprofitable and risky lending opportunities, shifting losses onto senior claimants. In contrast, Phillipon and Schnabl (2013) argue that the combination of preferred stock and warrants could be used to mitigate adverse selection problems while alleviating debt overhang. In another theoretical line of work, Bebchuk and Goldstein (2011) model credit freezes as coordination problems using the global games approach. In their model, government infusions of capital into banks will only jump start lending when a critical mass of worthy private sector projects will receive loans.

Empirically, Duchin and Sosyura (2012) find that banks increased the riskiness of the mortgage and business loans after taking TARP funds based on their analysis of micro level loan data. This is consistent with the predictions of Wilson (2009a) and Wilson and Wu (2010) that the preferred stock in TARP encourages riskier loan selection because it is a senior security to common equity. Black and Hazelwood (2012) find that large banks are significantly more likely to increase their holdings of risky commercial loans after taking TARP. Yet they find that smaller banks actually decrease their commercial loan origination. Wilson (2013) and Georgieva and Wilson (2010) indicate that a high percentage of small banks fell behind on their TARP dividends within the first 2 years. These distressed banks may have been restrained by regulators from originating risky loans.

Bayazitova and Shivdasani (2012), Duchin and Sosyura (2013), Jordan et al. (2011), Liu et al. (2011) and Ng et al. (2010) all look at the characteristics of banks selected for capital infusions into the TARP. Banks that were politically connected and had lower price to book ratios and greater asset quality were more likely to receive TARP funds. Li (2011) and Taliaferro (2009) find that TARP funds were associated with greater lending for TARP recipients, but the effects were modest. Taliaferro (2009) finds that most of the TARP funds were used to plug capital holes.

Wilson and Wu (2012) and Bayazitova and Shivdasani (2012) find that, after controlling for other factors, banks with higher paid CEOs were more likely to exit TARP. Cadman et al. (2010) outline the executive pay restrictions associated with taking TARP funds. These restrictions come from the TARP legislation (EESA), US Treasury regulations and the stimulus bill – the American Recovery and Reinvestment Act of 2009 (ARRA). They include limits on the tax deductibility of bonuses, limits on 'luxury' expenditures, non-binding say on pay votes, limits on golden parachutes for the most highly paid executives, and restrictions on the payments of cash bonuses and the early vesting of those bonuses. These restrictions provided an extra motivation for banks' executives to exit the program early and discouraged many banks from entering the program, especially after the ARRA legislation passed on 17 February 2009. In addition, banks within the TARP had significantly more turnover of their CEOs according to Cazier (2012). Kim

(2010) finds that TARP banks' share prices experienced negative and significant abnormal returns after more executive pay restrictions were announced. This effect was strongest for the largest banks. Thus, these executive pay regulations may have been perceived as hurting the common shareholders of TARP banks.

Overall the TARP probably helped the senior creditors of TARP banks more than their common stockholders. While Kim (2010) and Elyasiani et al. (2011) do find positive abnormal returns associated with the announcement of the Capital Purchase Program investments, Veronesi and Zingales (2010) find little benefit for the shareholders of the first banks included in program. Instead, Veronesi and Zingales (2010) estimate that only bondholders and preferred stockholders benefited from the TARP monies and the simultaneous announcement of a Federal Deposit Insurance Corporation (FDIC) guarantee for senior unsecured debt. Stock and Kim (2013) find that preferred stockholders with senior claims to the US Treasury experienced significantly positive abnormal returns after the banks were granted TARP capital injections.

Taxpayers have made money on the Capital Purchase Program (CPP) overall. $191.1 billion had been repaid by 30 June 2012, according to SIGTARP (2012, pp. 46). The CPP also realised dividends, interests, capital gains and proceeds from warrant sales of $26.2 billion. Thus, repayments and proceeds from the CPP topped $217.3 billion. That guarantees a nominal profit on this program of $12.4 billion with many more investments outstanding.

Khan and Dushyantkumar (2012) and Wilson andWu (2012) find that banks repaying TARP are significantly more likely to issue new stock as part of a Seasoned Equity Offering (SEO). Thus, TARP repayments often coincided with raising private capital. Cornett et al. (2013) finds that repayments are positively correlated with improvements in operating performance since taking TARP funds.

Taxpayers have received $11.7 billion through mid-2012 from payments of dividends and interest from the CPP investments. While the largest banks have generally paid their dividends and repaid TARP early, according to Georgieva and Wilson (2010) and Wilson and Wu (2012), smaller banks have been significantly more likely to fall behind on their TARP investments. While dividend payments have made the CPP a profitable program, hundreds of TARP recipients have skipped taxpayer dividends. According to Wilson (2013), the 1-year dividend skipping rate of the typical TARP bank indicates that most banks' TARP preferred stock should be rated as junk – below investment grade. That study finds that banks with lower capital ratios, more allowances for loan losses, and more net charge offs are significantly more likely to miss their bailout dividends.

The TARP legislation required that the US Treasury obtain warrants in companies benefiting from TARP investments. Warrants are rights to buy newly issued shares at a preset exercise price. Warrants allow their owners to benefit from increases in the companies' share prices above the exercise price. In practice, for publicly traded companies warrants behave similarly to call options. Warrant sales have accounted for $7.7 billion in proceeds from the TARP, according to SIGTARP (2012, p. 88). Wilson (2009b) explains the features and valuation of the warrants using the Goldman Sachs TARP warrants as an example. (Warren Buffet's investment of preferred stock and warrants in Goldman Sachs was very similar to the CPP.Moreover, Mr. Buffet was credited by Paulson (2010, p. 355) as suggesting the ultimate dividend rate of the CPP investments.) Wilson (2009c) argues that the first negotiated warrant repurchase price was too low. Wilson (2012) finds that low offers in warrant negotiations were associated with significantly lower negotiated purchase prices as a percent of estimated values by the Treasury's external experts and the Congressional Oversight Panel. Puente (2012), in a small sample, finds no significant evidence of political influence playing a role in the warrant negotiations between the banks repaying TARP funds and the US Treasury. Wilson (2010a) argues that the warrant auctions substantially increased the size of the publicly traded warrant market in the USA. Thus, these new issues of publicly traded warrants as a result of

TARP may open new avenues of empirical research into warrant securities performance and pricing.

The numbers on repayments are inflated by about $3.1 billion due to refinancing of TARP capital into other government programs according to SIGTARP (2012, pp. 87–88). Many small banks refinanced the TARP CPP with US Treasury capital that had more generous terms. Over a hundred small banks refinanced into the TARP's Community Development Capital Initiative (CDCI) or the non-TARP Small Business Lending Fund (SBLF). Wilson (2011b) argues that selection in the former program was more likely for credit unions headquartered in influential Congress members' districts. Both programs gave banks the opportunity to pay much lower dividends than in the CPP.

Should We Be Troubled That the TARP Bought So Few Troubled Assets?

The monies ultimately allocated to buying troubled assets (mortgage-related assets) in the TARP were actually very small. The largest 'toxic asset' program in the TARP was the Public Private Investment Partnership (PPIP). It will spend no more than $21.9 billion of taxpayer monies purchasing mortgage-related assets. This is far less than the $500–$750 billion in mortgage purchases which were advertised when Secretary Timothy Geithner rolled out this program in March 2009. The PPIP has had taxpayers contributing 75% of the purchase price of commercial and residential mortgage securities. Two thirds of the taxpayer monies were in the form of low interest rate, non-recourse debt. Taxpayers also contributed half the equity to the PPIP investment funds. Thus, two-third of taxpayers' funds were non-recourse loans to the PPIP funds and one-third was an equity stake in the funds. Nine original private asset managers controlled the investments of these funds and contributed half of their equity. The PPIP funds only invested in residential and commercial mortgage backed securities. They did not invest in whole loans. (Originally, the TARP was going to help fund purchases of troubled legacy loans from open banks with the help of the FDIC, but the FDIC chose to back out of those plans according to Wilson (2010c).) The debt stake received a very low interest rate tied to the LIBOR. The equity stakes of taxpayers and private investors received any residual profits from the toxic asset purchases and sales.

Originally, the idea of buying toxic assets was to clean up banks' balance sheets. Yet banks with troubled assets may not be as eager to part with them as many policymakers assumed.Wilson (2010b) argues that banks will value nonperforming assets more highly than market participants because bank shareholders benefit from the volatility of those toxic loans. With limited liability and high leverage, unloading risky assets for safe assets would decrease the value of bank shareholders' limited liability option. Policy makers seemed to understand this difficulty when designing the PPIP. Wilson (2011a) explains how the non-recourse loans extended by taxpayers to PPIP asset managers gave those managers option-like incentives to bid aggressively for residential and commercial mortgages. Bhansali and Wise (2009) argue that the government's non-recourse loans are less risky because TALF managers hold a portfolio of securitised assets. The risk of the portfolio is less than the weighted average risk of its component assets.

So far there has been no empirical work on the PPIP program because the details of the assets purchased are not available. This is likely to change after the 3-year investment period ends in 2012 and 2013 or when the TARP loans expire between 2018 and 2020. After that time, it is likely that more data on the investments will become available to researchers. It would also be interesting to study the characteristics of the asset managers who were ultimately chosen to run the TARP program.

The TARP also contributed $4.3 billion to help finance the purchase of a variety of securitised assets through the Term Asset-Backed Liquidity Facility (TALF) run by the Federal Reserve. Yet most of this commitment has been returned by mid-2012. The TARP monies in TALF were put in a subordinated position to the Federal Reserve's

loans. The TARP stakes were used to cover losses if any asset managers borrowing from the Federal Reserve defaulted on loans that were used to purchase securitised commercial mortgages, student loan receivables and auto loan receivables. Wilson (2011e) finds that the subsidy rate for this program's loans to purchase commercial mortgage-backed securities (CMBS) was 34% at origination. Yet recoveries in CMBS prices encouraged many asset mangers to repay the Federal Reserve and the Treasury by the third quarter of 2010 when the subsidy for the taxpayer financing turned negative. In addition, the largest asset managers were the quickest to repay the government loans to buy CMBS.

Did the Auto Bailout Rewrite the Bankruptcy Code?

SIGTARP (2012, pp. 38–40) reports that taxpayers spent $79.7 billion on the bailouts of General Motors (GM) and Chrysler. At the time of writing, taxpayers had fully exited their investment in Chrysler, realising a loss of $2.93 billion on an investment of $5.39 billion. (Chrysler received $4.6 billion in loan commitments that it did not use according to SIGTARP (2012, p. 145.) Roe and Skeel (2010) note that the government was prepared to fund Chrysler's assets worth $2 billion with $15 billion in loans.) Yet taxpayers have a common stock stake and cash proceeds from stock sales that at the time of writing are worth far less than the $49.5 billion investment in GM. Moreover, the current stock price of GM indicates that taxpayers are likely to take large losses on the GM investment. The former financing arm of GM, Ally Bank, which was formerly called the General Motors Acceptance Corporation (GMAC), received most of the balance of TARP funds ($17.2 billion) and is still privately held with taxpayers holding a 72% stake in the lender.

The bailouts of the automakers seem to have little to do with the real estate crisis. They involved cyclical industrial companies whose eventual bank-ruptcy barely created a ripple in financial markets. By all accounts, both companies had a failing business model and high cost structure. Roe and Skeel (2010) write, 'Chrysler was a weak producer, making cars that had limited consumer acceptance, in an industry suffering from substantial domestic and worldwide overcapacity'. Ramseyer and Rasmussen (2011) are even less optimistic about GM's business prospects when recounting how it accumulated net operating losses of $45 billion leading up to its bankruptcy:

> Year after year, General Motors lost money—enormous sums of money. It designed cars. It built cars. But no one wanted to buy the cars. Over time, it accumulated huge operating losses ('net operating losses', or NOLs). The tax code let GM carry forward these NOLs into the future. It let the firm save the losses for that day in the future when it would once again sell cars that people wanted.
>
> The day never came.

Chapter 11 reorganisation was probably necessary to improve both companies' competitiveness. Yet taxpayers appear to have paid a very high price for postponing the companies' bankruptcies by a few months.

The returns from the auto bailouts of GM and Ally Financial are significantly overstated. At least two of GM's analysts estimated the present value of its $45 billion of NOLs at $13 billion, according to Ramseyer and Rasmusen (2011). NOLs cannot be transferred when existing owners of the corporation sell greater than a 50% stake to new owners in less than a 3 year period. In the GM reorganisation, taxpayers received a 62% ownership stake in the post-bankruptcy GM.Without IRS notice 2010–2, which appears to have arbitrarily contradicted years of tax policy, the already unprofitable GM bailout would look even uglier.

Ramseyer and Rasmusen (2011) assert that Citigroup and AIG also benefited from this favourable tax treatment of NOLs for US Treasury-owned companies. These firms had NOLs of $46.1 billion and $34.9 billion, respectively. The greatest taxpayer stake in Citigroup was 34% of the lender's common stock. The $25 billion Citigroup investment which was converted into common stock generated the largest profit of any bank investment in the TARP, $6.85 billion according to Wilson (2011d). Ramseyer and

Rasmusen (2011) consider the raise of an additional $24 billion in common equity in Citigroup sufficient to exceed the 50% threshold. At the time of writing, the US Treasury and Federal Reserve have already booked a profit in the $182 billion AIG bailout with a 22% government stake in AIG remaining, according to Dennis (2012). The taxpayer stake in AIG was as high as 92%. Thus, it would have triggered the prohibitions on NOL transfer without a special ruling from the IRS. Thus, any nominal profits in the AIG and Citigroup investments would have to be weighed against the lost taxes from waiving the NOL rules for these firms.

Little empirical work has been done on the auto bailouts. The dealer closings and plant closing decisions, which came out of the bankruptcy, may be routes for future study. Most scholarship is from legal scholars, whose focus is mainly on the issues in the bankruptcies of GM and Chrysler. A few examples are Adler (2010), Ayotte and Skeel (2010) and Roe and Skeel (2010).

Fremeth et al. (2012) have attempted to get around the small sample problem by modelling a Chrysler that was not bailed out as a weighted average of similar car companies in 2009 based on a synthetic control approach, which is more common in management literature. They argue that the performance of Chrysler could be estimated from a weighted average of other car companies that did not go bankrupt in 2009. Fremeth et al. (2012) conclude that a Chrysler without a bailout would have produced 20% more cars. Other routes of analysis of the auto bailouts may be simulations such as those used to study the US Treasury's stock sales of Citigroup by Wilson (2011c).

The Slow Pace of Mortgage Modifications and Literature on the TARP's's Foreclosure Mitigation Efforts

One of the more popular aspects of the TARP program was the government's attempt to keep struggling homeowners in their houses. Foreclosures may represent social waste because homes lie vacant. Foreclosed homes are more likely to be poorly cared for and may degrade the quality of life in their neighbourhoods as well as increasing the inventory of homes up for sale. Yet the first TARP programs for foreclosure mitigation were not announced until after President Obama was inaugurated in February 2009. (The Bush Administration and the Paulson Treasury did sponsor a couple of non-TARP foreclosure mitigation programs – FHA-Secure and Hope for Homeowners – but the actual number of mortgage modifications coming from those efforts were less than 5000 loans according to Gans (2012).) After that these programs took a long time to gain momentum. Despite the TARP foreclosure mitigation efforts, RealtyTrac (2011) found that there were in excess of 3.8 million properties in foreclosure in the USA, a record number, by the end of 2010. When a property has mortgages with principal balances in excess of the home's value, the homeowner with negative equity is said to be 'underwater'. The Economist (2010) reported that roughly a quarter of US households were underwater on their mortgages and about 4 million households had mortgages for twice what their home was worth.

The US Treasury is certain to spend far less than the $50 billion that was initially allocated to the effort, despite the foreclosure mitigation programs being extended through 31 December 2013, at the time of writing. By midyear 2012, there were over half a dozen foreclosure mitigation programs within TARP under the umbrella of the Making Home Affordable (MHA). These programs, through 30 June 2012, have spent over $4.5 billion, according to SIGTARP (2012, p. 40).

By any measure, the largest foreclosure mitigation program in TARP is the Home Affordable Modification Program (HAMP). HAMP provides TARP funds to temporarily reduce home payments that exceed 31% of the borrowers' gross income. This program had by 30 June 2012, resulted in 1.04 million 'permanent' mortgage modifications according to SIGTARP (2012, p. 69). (The title 'permanent modification' is a bit of a misnomer, since permanent modifications begin to be phased out after 5 years.) Yet SIGTARP (2012, p. 69) points out that about

half of those permanent modifications were funded outside of TARP through the government-sponsored entities Fannie Mae and Freddie Mac. Moreover, this falls short of the administration's goal to modify the mortgages of four million homeowners. The US Treasury (2012a) said that the median borrower with a permanent modification has seen his or her monthly payment reduced by $538, or 38% of that person's monthly payment.

Empirical analysis of the HAMP and related foreclosure mitigation programs is still a largely untapped source of academic research into the TARP program. Many such studies seem possible because of the large universe of eligible loans and large number of modified mortgages. This contrasts with the seeming impossibility of empirical work on the AIG and auto bailouts because so few firms were actually eligible. Yet the author only knows of one empirical study on HAMP. Agarwal et al. (2012) use data from the office of the Comptroller of the Currency to estimate that the HAMP program reduced mortgage foreclosure rate for eligible borrowers by 0.48%. I have not seen any serious theoretical papers modelling the HAMP program, yet there is a literature on mortgage renegotiation that predates the current crisis according to Gerardi and Li (2010).

White (2009) argues that the cost of recent foreclosures was 55% of loan principal, whereas the cost of modification was less than 9% of the loan balance. This gap indicated to policy makers that foreclosure mitigation efforts may even be in investors' interests. Yet theoretical studies by Ambrose and Capone (1996), Riddiough and Wyatt (1994), Wang et al. (2002) and Foote et al. (2009) argue, respectively, that the chances that borrowers will become current without lender concessions, asymmetric information between borrower and lender, reputational concerns and falling house prices can make lenders and servicers wary of making concessions to struggling borrowers.

It is an open question whether legal obstacles have prevented some lender forbearance. Piskorski et al. (2010) and Agarwal et al. (2011) find that banks' portfolio loans are much more likely to be modified versus securitised loans, yet Adelino et al. (2009) see no evidence that securitised loans have lower modification rates. Adelino et al. (2009) find that in distressed loans originated in 2005 to 2007 the foreclosure rate was 10 times larger than the modification rate and these differences cannot be explained by legal obstacles to renegotiation of securitised mortgages. Instead, it is not optimal for investors to modify mortgages because there is a large 'self-cure' and 're-default risk'. The self-cure risk is the chance that the borrower would become current on his or her mortgage without a modification. There is also a chance that borrowers will re-default after a modification postponing the inevitable foreclosure in a declining housing market. Adelino et al. (2009) put these combined hazards at greater than 60%. Thus, conceivably over 60% of the time investors will lose out by modifying mortgages without government carrots.

Clearly more work can be done on the HAMP and other foreclosure mitigation programs in TARP. However, the small size of these efforts in dollar terms may discourage potential scholars in this area from learning the lessons from TARP's attempts to help underwater homeowners. These foreclosure mitigation programs represent pure subsidies, as opposed to the investments in the banks, AIG, the auto makers, CMBS and RMBS, where all or a large percentage of the money is repaid. Thus it would be very costly to for these programs to continue to be largely ignored by scholars.

The *ad hoc* Bailouts of AIG, Citigroup, and Bank of America

American International Group (AIG) received $67.8 billion from the TARP as part of a Treasury and Federal Reserve rescue that topped $182 billion. The full amount was returned prior to 2013, and the Federal Reserve and the US Treasury by the end of 2012 had closed out that investment with a $22.7 billion profit (U.S. Treasury 2012b). (TheTARP program may technically be in the red on AIG because much of the returns from the AIG investment come from a 79.9% stake from the

initial Federal Reserve loans to AIG.) While Sjostrom (2011) discusses the AIG bailout in detail, I have not seen any rigorous empirical or theoretical study of the AIG bailout. Unlike the banks in the CPP, there is only one company in the TARP program for AIG. It also is one of only a handful of insurance companies which received TARP funds. Thus the TARP bailouts of AIG afford little room for empirical work. Taxpayers still retained a significant minority common stock stake in AIG in October 2012 that may generate further returns.

Bank of America and Citigroup received an additional $20 billion each in exchange for preferred stock and warrants as part of the Targeted Investment Program (TIP). (They both ultimately received $25 billion in exchange for preferred stock and warrants from the CPP, too.) They also received sizable asset guarantees, for assets totalling $118 billion and $301 billion, respectively according to SIGTARP (2012, pp. 126–7). Both banks cancelled the asset guarantees and redeemed the TIP preferred stock prior to the end of 2009, to escape executive pay restrictions. Neither the government nor any scholar has circulated a study that seriously valued the huge asset guarantees offered to these banks. That would be an interesting exercise that we should see.

Conclusion

Most of the programs in the TARP, the so-called $700 billion bailout, have been largely ignored in terms of rigorous scholarly research. While the bank investment programs have received very extensive academic scrutiny, the troubled asset, auto and housing bailouts need more research. This was an unprecedented intervention by US taxpayers into financial markets. It generated great public scrutiny into the financial sector and created the political climate for the overhaul of the financial sector by the Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010. There is more room for empirical and theoretical research into the hundreds of billions of dollars spent on the troubled asset, auto, and housing bailouts in the TARP.

See Also

- ▶ Fall of AIG
- ▶ Fannie Mae, Freddie Mac and the Crisis in US Mortgage Finance
- ▶ Lehman Brothers Bankruptcy, What Lessons Can Be Drawn?
- ▶ Subprime Mortgage Crisis

Bibliography

Adelino, M., K. Gerardi, and P.S. Willen. 2009. *Why don't lenders renegotiate more home mortgages? Redefaults, self-cures, and securitization*, Federal reserve bank of Atlanta working paper 2009–17. Cambridge, MA: National Bureau of Economic Research.

Adler, B. 2010. Chapter 11 at the crossroads: Does reorganization need reform?: A reassessment of bankruptcy reorganization after Chrysler and general motors. *American Bankruptcy Institute Law Review* 18(2): 305–404.

Agarwal, S., G. Amromin, I. Ben-David, S. Chomsisengphet, and D.D. Evanoff. 2011. The role of securitization in mortgage renegotiation. *Journal of Financial Economics* 102(3): 559–578.

Agarwal, S., G. Amromin, I. Ben-David, S. Chomsisengphet, T. Piskorski, and A. Seru. 2012. *Policy intervention in debt renegotiation: Evidence from the Home Affordable Modification Program*. Available at: http://ssrn.com/abstract=2138314. Accessed 18 Jan 2013.

Ambrose, B., and C. Capone. 1996. Cost-benefit analysis of single-family foreclosure alternatives. *Journal of Real Estate Finance and Economics* 13(2): 105–120.

Ayotte, K., and D.A. Skeel. 2010. Bankruptcy or bailouts? *Journal of Corporation Law* 35(3): 469–498.

Bhansali, V., and M.B. Wise. 2009. How valuable are the TALF puts? *Journal of Fixed Income* 19(2): 71–75.

Bayazitova, D., and A. Shivdasani. 2012. Assessing TARP. *Review of Financial Studies* 25(2): 377–407.

Bebchuk, L., and I. Goldstein. 2011. Self-fulfilling credit market freezes. *Review of Financial Studies* 24(11): 3519–3555.

Black, L., and L. Hazelwood. 2012. *The effect of TARP on bank risk-taking*, Board of governors of the federal reserve system international finance discussion paper, IFDP 1043. Washington, D.C: Federal Reserve Board.

Cadman, B., Carter, M. E. and Lynch, L. J. 2010. *Executive pay restrictions: Do they restrict firms' willingness to participate in TARP?*. Working paper. David Eccles School of Business. University of Utah.

Cazier, R. A. 2012. *Executive compensation and retention outcomes associated with TARP participation*. SSRN working paper. Available at: http://ssrn.com/abstract=2142909. Accessed 18 Oct 2012.

Cornett, M. M., L. Li, and H. Tehranian. 2013. The performance of banks around the receipt and repayment of TARP funds: over-achievers versus under-achievers. *Journal of Banking and Finance* (forthcoming).

Dennis, B. 2012. AIG bailout success a two-sided coin. *Washington Post*, 10 September. Available at: http://www.washingtonpost.com/business/economy/aigbailout-success-a-two-sided-coin/2012/09/10/4131b8f0-fb5f-11e1-b153-218509a954e1_story.html. Accessed 30 Sept 2012.

Duchin, R. and D. Sosyura. 2012. *Safer ratios, riskier portfolios: Banks' response to government aid*. Finance department working paper. University of Michigan.

Duchin, R., and D. Sosyura. 2013. The politics of government investment. *Journal of Financial Economics* (forthcoming).

Economist, The. 2010. *Drowning or waiving, the policy options for alleviating America's huge negative-equity problem*, 21 October. Available at: http://www.economist.com/node/17305544. Accessed 24 Sept 2012.

Elyasiani, E., L. J. Mester, and M. S. Pagano. 2011. *Large capital infusions, investor reactions, and the return and risk performance of financial institutions over the business cycle and recent financial crisis*. FRB of Philadelphia working paper No. 11–46. Available at: http://ssrn.com/abstract=1942323. Accessed 18 Oct 2012.

Foote, C., K. Gerardi, L. Goette, and P. Willen. 2009. Reducing foreclosures: No easy answers. *NBER Macroeconomics Annual* 24: 89–138.

Fremeth, A., G. L. F. Holburn, and B. K. Richter. 2012. *Did Chrysler benefit from government assistance? Making causal inferences in small samples using synthetic control methodology*. SSRN working paper. Available at: http://ssrn.com/abstract=2135294. Accessed 18 Oct 2012.

Gans, M. 2012. HAMP: Doomed from the start. *Cornell Real Estate Review* 10(1): 54–71.

Gerardi, K., and W. Li. 2010. Mortgage foreclosure prevention efforts. *Economic Review, Federal Reserve Bank of Atlanta*, 95. Available at: http://hdl.handle.net/10419/57666. Accessed 1 Oct 2012.

Georgieva, D., and L. Wilson. 2010. *TARP's dividend skippers*. SSRN working paper. Available at: http://ssrn.com/abstract=1654677. Accessed 14 Aug 2010.

Green, M. N., and K. Hudak. 2009. Congress and the bailout: explaining the bailout votes and their electoral effect. *American Political Science Association. Legislative Studies Section Newsletter*, 32(1). Available at: http://www.apsanet.org/_lss/Newsletter/jan2009/GreenHudak.pdf. Accessed 18 Oct 2012.

Jordan, D.J., D. Rice, J. Sanchez, and D.H. Wort. 2011. Explaining bank marketto-book ratios: Evidence from 2006 to 2009. *Journal of Banking and Finance* 35(8): 2047–2055.

Khan, M. and V. Dushyantkumar. 2012. *Bank recapitalization through SEOs: Why is this time different?* SSRN working paper. Available at: http://ssrn.com/abstract=1971531. Accessed 18 Oct 2012.

Kim, W. Y. 2010. *Market reaction to limiting executive compensation: Evidence from TARP firms*. SSRN working paper. Available at: http://ssrn.com/abstract=1553394. Accessed 14 Oct 2012.

Li, L. 2011. *TARP funds distribution and bank loan growth*. SSRN working paper. Available at: http://ssrn.com/abstract=1515349. Accessed 17 Oct 2012.

Liu, W., J. W. Kolari, T. K. Tippens, and D. R. Fraser. 2011. *Did capital infusions enhance bank recovery from the Great Recession?* SSRN working paper. Available at: http://ssrn.com/abstract=1962795. Accessed 14 Oct 2012.

Ng, J., F. P. Vasvari, R. Wittenberg-Moerman. 2010. *Were healthy banks chosen in the TARP capital purchase program?* SSRN working paper. Available at: http://ssrn.com/abstract=1566284. Accessed 22 May 2010.

Paulson, H.M. 2010. *On the brink: Inside the race to stop the collapse of the global financial system*. New York: Business Plus.

Philippon, T., and P. Schnabl. 2013. Efficient recapitalization. *Journal of Finance* (forthcoming).

Piskorski, T., A. Seru, and V. Vig. 2010. Securitization and distressed loan renegotiation: Evidence from the subprime mortgage crisis. *Journal of Financial Economics* 97(3): 369–397.

Propublica.org. 2013 *Bailout timeline: Another day, another bailout*. Available at: http://projects.propublica.org/bailout/main/timeline. Accessed 10 Jan 2013.

Puente, L. 2012. Political influence and TARP: An analysis of Treasury's disposition of CPP warrants. *PS: Political Science Politics* 45(2): 211–217.

Ramirez, C.D. 2011. The $700 billion bailout: A public-choice interpretation. *Review of Law Economics* 7(1): 291–318.

Ramseyer, J.M., and E. Rasmussen. 2011. Can the Treasury exempt companies it owns from taxes? The $45 billion general motors net operating loss carryforward. *The Cato Papers on Public Policy* 1(1): 1–54.

RealtyTrac.com (2011) *Record 2.9 Million U.S. properties receive foreclosure filings in 2010 despite 30-Month low in december*. Available at: http://www.realtytrac.com/content/press-releases/record-29-million-us-properties-receive-foreclosure-filingsin-2010-despite-30-month-low-in-december-6309. Accessed 24 Sept.

Riddiough, T.J., and S.B. Wyatt. 1994. Wimp or tough guy: Sequential default risk and signaling with mortgages. *Journal of Real Estate Finance and Economics* 9(3): 299–321.

Roe, M.J., and D.A. Skeel. 2010. Assessing the Chrysler bankruptcy. *Michigan Law Review* 108(5): 727–772.

SIGTARP. 2012. *Quarterly report to congress, July 25, 2012*. Office of the Special Inspector General for the Troubled Asset Relief Program. Available at: http://www.sigtarp.gov/Quarterly%20Reports/July_25_2012_Report_to_Congress.pdf. Accessed 15 Oct 2012.

Sjostrom, W. K., Jr. 2011. The fall of AIG. In *The new Palgrave dictionary of economics*. Online Edition, eds. Steven N. Durlauf and Lawrence E. Blume.

Stock, D., and D. H. Kim. 2013. Impact of the TARP financing choice on existing preferred stock. *Journal of Corporate Finance* (forthcoming). Available at: http://ssrn.com/abstract=2120936. Accessed 18 Oct 2012.

Taliaferro, R. 2009. *How do banks use bailout money? Optimal capital structure, new equity, and the TARP*. SSRN working paper. Available at: http://ssrn.com/abstract=1481256. Accessed 2 Aug 2010.

U.S. Treasury. 2012a. *Making home affordable program performance report through July 2012*. Available at: http://www.treasury.gov/initiatives/financial-stability/reports/Documents/July%202012%20MHA%20Report_SERVICER%20ASSESSMENTS_Final.pdf. Accessed 18 Oct 2012.

U.S. Treasury. 2012b. *Press release: Treasury sells final shares of AIG common stock, positive return on overall AIG commitment reaches $22.7 billion*. Available at: http://www.treasury.gov/press-center/press-releases/Pages/tg1796.aspx. Accessed 12 Jan 2013.

Veronesi, P., and L. Zingales. 2010. Paulson's gift. *Journal of Financial Economics* 97(3): 339–368.

Wang, K., L. Young, and Y. Zhou. 2002. Non-discriminating foreclosure and voluntary liquidating costs. *Review of Financial Studies* 15(3): 959–985.

Wen, H. J. 2011. *The political economy of TARP bank bailouts*. Masters of public policy thesis, Georgetown University. Available at: http://repository.library.georgetown.edu/bitstream/handle/10822/553957/wenHua.pdf. Accessed 12 Jan 2013.

White, A.M. 2009. Deleveraging the American homeowner: The failure of 2008 voluntary mortgage contract modifications. *Connecticut Law Review* 41: 1107.

Wilson, L. 2009a. *Debt overhang and bank bailouts*. SSRN working paper. Available at: http://ssrn.com/abstract=1336288. Accessed 22 May 2009.

Wilson, L. 2009b. The Goldman Sachs warrants. *Review of Business*, 30(1): 4–32.

Wilson, L. 2009c. *Valuing the first negotiated repurchase of the TARP warrants*. SSRN working paper. Available at: http://ssrn.com/abstract=1404069. Accessed 17 Oct 2012.

Wilson, L. 2010a. The biggest warrant auction in U.S. history. *Research in Business and Economics Journal* 3: 1–12.

Wilson, L. 2010b. The put problem with buying toxic assets. *Applied Financial Economics* 20(1): 31–35.

Wilson, L. 2010c. Slicing the toxic pizza: An analysis of FDIC's legacy loan program for receivership assets. *International Journal of Monetary Economics and Finance* 3(3): 300–309.

Wilson, L. 2011a. A binomial model of Geithner's toxic asset plan. *Journal of Economics and Business* 63(5): 349–371.

Wilson, L. 2011b. *Political influence and TARP investments in credit unions*. SSRN working paper. Available at: http://ssrn.com/abstract=1698945. Accessed 17 Oct 2012.

Wilson, L. 2011b. Selling Citigroup: A simulation of the U.S. Treasury's $37 billion TARP share sale. *Review of Business* 31(2): 3–14.

Wilson, L. 2011c. Stocks demand curves and TARP returns. *Journal of Financial Economic Policy* 3(3): 229–242.

Wilson, L. 2011e. *Toxic asset subsidies and the early redemption of TALF loans*. Available at: http://ssrn.com/abstract=1742640. Accessed 17 Oct 2012.

Wilson, L. 2012. Anchoring bias in TARP warrant negotiations. *Journal of Economics and Business* 64(1): 63–76.

Wilson, L. 2013. TARP's deadbeat banks. *Review of Quantitative Finance and Accounting* (forthcoming).

Wilson, L., and W.Y. Wu. 2010. Common (stock) sense about risk-shifting and bank bailouts. *Financial Markets and Portfolio Management* 24(1): 3–29.

Wilson, L., and Y. Wu. 2011. Overpaid CEOs got FDIC debt guarantees. *SSRN Working Paper*. Available at: http://ssrn.com/abstract=1977345. Accessed 17 Oct 2012

Wilson, L., and W.Y. Wu. 2012. Escaping TARP. *Journal of Financial Stability* 8(1): 32–42.

Wurtz, K.P. 2010. Fishing for TARP: constituency service, financial firms, and the political determinants and consequences of the 2008 TARP votes. *Working Paper*, Department of International Relations, Lehigh University. Available at: http://kwurtz.files.wordpress.com/2010/02/wurtz-apsa-2010.pdf. Accessed 19 Oct 2012.

Trust in Experiments

Iris Bohnet

Abstract

Trust is the willingness to make oneself vulnerable to another person's actions, based on beliefs about that person's trustworthiness. This article focuses on interpersonal trust and trustworthiness between two people, a trustor and a trustee, as measured in laboratory experiments. A trustee behaves trustworthily if he voluntarily refrains from taking advantage of the trustor's vulnerability. Trust applies to all transactions where the outcome is partly under the control of another person and not fully contractible. The article discusses measurement issues, the motives for and influences on trust and trustworthiness (incentives, repetition

and demographic variables) and questions of external validity.

Keywords

Cooperation; Fairness; Folk theorem; Investment game; Public goods games; Reciprocity; Repeated games; Social preferences; Trust; Trust in experiments; Trust games; Ultimatum games

JEL Classifications

C9

Trust is the willingness to make oneself vulnerable to another person's actions, based on beliefs about that person's trustworthiness.

This article focuses on interpersonal trust between two people, a trustor and a trustee. A trustee behaves trustworthily if he voluntarily refrains from taking advantage of the trustor's vulnerability. Trust applies to all transactions where the outcome is partly under the control of another person and not fully contractible, for example, between employers and employees or patients and doctors. Trust and trustworthiness are typically measured in surveys or laboratory experiments. We shortly discuss some survey evidence but focus on behavioural measures of trust.

Measurement

Trust attitudes have typically been measured by the following survey question (for example, used in the General Social Survey and the World Values Survey): 'Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?' Based on this question, trust has declined dramatically across the world since the 1960s. There are noticeable cross-country differences, with Scandinavians most and Latin Americans least likely to trust others. An empirical literature building on Knack and Keefer (1997) shows that trust attitudes are positively correlated with various measures of a country's economic performance.

Around 1990, two seminal papers started a new wave in the economic research on trust. In 1988 Camerer and Weigelt employed a binary-choice trust game, and in Berg et al. 1995 the 'investment game' to study trust. In the binary-choice trust game, the trustor decides between a sure outcome and trust. If she chooses the sure thing, she and her trustee both receive (S, S). If she is willing to trust, both either end up with a moderate payoff exceeding S (M, M), or the trustor receives a lower payoff than if she had not trusted, and the trustee the highest possibly payoff, (L, H). Thus, for the trustor, $M > S > L$, and for the trustee, $H > M > S$. In the investment game, a trustor and a trustee are endowed with a certain amount of money, A (in some experiments, only trustors are endowed). The trustor can send any amount, $X \leq A$, to the trustee. X is multiplied by $k > 1$ by the experimenter. In most experiments, $k = 3$. Trustees receive kX and then decide how much of it, $Y \leq A + kX$, to return to their trustor. The final payoffs are $A - X + Y$ for the trustor and $A + kX - Y$ for the trustee. X is commonly referred to as trust and Y, or more precisely, Y/X, measures trustworthiness for $X > 0$. In both games, the equilibrium prediction based on selfish money-maximization and rationality is zero trustworthiness and zero trust.

The relationship between trust attitudes, as measured in surveys, and trust behaviour, as measured in experiments, is not clear. Some have found that they are related (for example, Fehr and Schmidt 2002), others that they are not (for example, Glaeser et al. 2000). While the investment game and the binary-choice trust game have turned out to be the most widely used games to study trust experimentally, related games include the 'gift exchange game', the 'moonlighting game,' and standard public goods games (for a review, see Camerer 2003).

What Motivates Trust and Trustworthiness?

Trust is based on preferences, namely, the willingness to be vulnerable to someone else, and on expectations, namely, the belief about someone

else's trustworthiness. A person's willingness to be vulnerable may be related to her attitudes to risk (for example, Eckel and Wilson 2004), her social preferences (for example, Cox 2004), and her willingness to accept the risk of betrayal (Bohnet and Zeckhauser 2004). Bohnet and Zeckhauser introduced an analytical framework to disentangle the various motives, and show that people dislike making themselves vulnerable to the actions of another person more than to natural circumstances. This suggests betrayal aversion: people care not only about outcomes but also about how outcomes come to be. This finding was supported by neuroscientific evidence (Kosfeld et al. 2005).

The relevance of expectations of trustworthiness for trust has typically been measured by including a question about trustors' beliefs. While this measure is not perfect, generally the relationship between expectations of trustworthiness and trust is very strong. For example, using a within-subject design with behavioural controls for risk and social preferences, Ashraf et al. (2006) found that expectations of trustworthiness explain most of the variance in trust in an investment game but that social preferences also matter.

Trustworthiness is based on trustees' social preferences, which may be related either to outcomes (for a survey, see Fehr and Schmidt 2002) or to what the trustors' actions reveal about their intentions. In a seminal paper, Rabin (1993) introduced a theoretical model of intention-based preferences, reciprocity, into the literature. A large number of empirical studies suggests the importance of reciprocity in trust interactions (for example, Fehr et al. 1997) although outcome-based social preferences also play an important role for trustworthiness (for example, Cox 2004; Ashraf et al. 2006).

What Influences Trust and Trustworthiness?

Incentives

According to most models, trustors should be more likely to trust the higher the expected returns are from trusting. Bohnet et al. (2006) measured the elasticity of trust and found that trust is responsive both to changes in the likelihood and to the cost of betrayal in Western countries. However, this does not necessarily apply in other parts of the world. For example, in Persian Gulf countries people hardly responded to such changes. Instead, many basically demanded a guarantee of trustworthiness before trusting, suggesting substantial aversion to betrayal. In addition, incentives may also not work as predicted by theory if they not only affect behaviour directly but also exhibit an influence on preferences, thus either fostering or undermining people's willingness to accept vulnerability and be trustworthy voluntarily (Bohnet et al. 2001).

Repetition

Generally, people are more likely to trust and be trustworthy in repeated than in one-shot interactions. Theoretically, this result is expected in a traditional model when interactions are indefinitely repeated (folk theorem) but not in finitely repeated games. In support of the theory, experimental evidence suggests that trust and trustworthiness rates are generally higher in indefinitely than in finitely repeated games but they are also higher in the latter than in one-shot interactions. The equilibrium prediction of no trust and trustworthiness is generally refuted, although trust and trustworthiness rates typically drop substantially as the end of the game draws nearer (for example, Gächter and Falk 2002).

Demographic Variables

Generally, the evidence is not as conclusive as we might expect or wish. While in theory variables such as gender, race or country of origin should be easy to control for, experiments produce different results precisely because of the different sets of control variables and the different subject pools used. The most promising approaches include those identifying overarching frameworks able to account for a variety of studies. We discuss three such frameworks here: history of discrimination, societal organization and market integration.

Groups that historically have been discriminated against, such as women and minorities, are generally less likely to trust. At the same time,

often these groups are more trustworthy (for example, Alesina and LaFerrara 2002; Buchan et al. 2003; Eckel and Wilson 2003).

Group-based societal organization based on long-standing relationships and repeated interactions within groups can substantially reduce the social uncertainty involved in trust. It is often referred to as 'collectivist' in contrast to the Western ' individualist' model of organization, which produces trust through more anonymous, institutional arrangements such as contracts and insurance. Trust in strangers has often been found to be higher in individualist (for example, the United States or Switzerland) than in collectivist countries (for example, Japan or the Persian Gulf countries), although the rather small number of studies and sample sizes does not allow any definite conclusions at this point (for example, Bohnet et al. 2006; but see also Croson and Buchan 1999).

The degree of market integration is related to norms of cooperation and fairness in public goods and ultimatum games. Similarly, the norms of reciprocity typically found in trust experiments in developed countries seem to apply more strongly in societies in which goods and services are exchanged in the market rather than in informal reciprocal-exchange arrangements. Greig and Bohnet's survey of the evidence Greig and Bohnet (2006) suggested that the positive relationship between trust and trustworthiness, normally taken to indicate reciprocity, is more pronounced in developed than in developing countries.

External Validity

Experiments allow for maximum internal control. Concerns typically arising in field settings such as lack of randomization, selection and endogeneity can easily be addressed by experimental design. To address concerns about the subject pools experimentalists typically use, that is, North American or European students, experiments are now run with representative samples (for example, Fehr et al. 2002 in Germany) and with student and non-student subjects in other parts of the world (for example, Cardenas and Carpenter 2005, for a survey). To directly test the external validity of trust experiments, Karlan (2005) ran investment games with members of a group lending association in Peru, and compared trustworthiness in the experiment with repayment rates. The more trustworthy subjects indeed were significantly more likely to repay their loans a year later.

See Also

- ▶ Altruism in Experiments
- ▶ Behavioural Game Theory
- ▶ Experimental Economics
- ▶ Public Goods Experiments
- ▶ Reciprocity and Collective Action
- ▶ Risk Aversion
- ▶ Social Capital

Bibliography

Alesina, A., and E. LaFerrara. 2002. Who trusts others? *Journal of Public Economics* 85: 207–234.

Ashraf, N., I. Bohnet, and N. Piankov. 2006. Decomposing trust and trustworthiness. *Experimental Economics* 9: 193–208.

Berg, J., J. Dickhaut, and K.A. McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 290–307.

Bohnet, I., B.S. Frey, and S. Huck. 2001. More order with less law: On contract enforcement, trust and crowding. *American Political Science Review* 95: 131–144.

Bohnet, I., B. Herrmann, and R. Zeckhauser. 2006. The requirements for trust in Gulf and Western countries. Working paper.

Bohnet, I., and R. Zeckhauser. 2004. Trust, risk and betrayal. *Journal of Economic Behavior and Organization* 55: 467–484.

Buchan, N., R. Croson, and S. Solnick. 2003. Trust and gender: An examination of behavior, biases, and beliefs in the investment game. Working paper, Wharton School, University of Pennsylvania.

Camerer, C.F. 2003. *Behavioral game theory.* Princeton: Princeton University Press.

Camerer, C.F., and K. Weigelt. 1988. Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56: 1–36.

Cardenas, J.C., and J. Carpenter. 2005. Experiments and economic development: Lessons from field labs in the developing world. Working paper, Middlebury College.

Cox, J.C. 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46: 260–281.

Croson, R., and N. Buchan. 1999. Gender and culture: International experimental evidence from trust games. *American Economic Review* 89: 386–391.
Eckel, C.C. and R.K. Wilson. 2003. Conditional trust: Sex, race and facial expressions in a trust game. Working paper, Rice University.
Eckel, C.C., and R.K. Wilson. 2004. Is trust a risky decision? *Journal of Economic Behavior and Organization* 55: 447–466.
Fehr, E., and K. Schmidt. 2002. Theories of fairness and reciprocity – Evidence and economic applications. In *Advances in economics and econometrics*, ed. M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge, MA: Cambridge University Press.
Fehr, E., S. Gächter, and G. Kirchsteiger. 1997. Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica* 64: 833–860.
Fehr, E., U. Fischbacher, B. von Rosenbladt, J. Schupp, and G. Wagner. 2002. A nation-wide laboratory-examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Schmollers Jahrbuch* 122: 519–542.
Gächter, S., and A. Falk. 2002. Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics* 104: 1–26.
Glaeser, E.L., D.I. Laibson, J.A. Scheinkman, and C.L. Soutter. 2000. Measuring trust. *Quarterly Journal of Economics* 115: 811–846.
Greig, F., and I. Bohnet. 2006. Is there reciprocity in a reciprocal-exchange economy? Evidence of gendered norms from a slum in Nairobi, Kenya. Working paper, Kennedy School of Government, Harvard University.
Karlan, D. 2005. Using experimental economics to measure social capital and predict financial decisions. *American Economic Review* 95: 1688–1699.
Knack, S., and P. Keefer. 1997. Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics* 112: 1251–1288.
Kosfeld, M., M. Heinrichs, P.J. Zak, U. Fischbacher, and E. Fehr. 2005. Oxytocin increases trust in humans. *Nature* 435: 673–676.
Rabin, M. 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281–1302.

Tsuru, Shigeto (1912–2006)

Tsuneo Nakauchi

Keywords

Business cycles; Japan, economics in; Marxian methodology; Tsuru, S

JEL Classifications

B31

Tsuru was born on 6 March 1912 in Oita and brought up in Nagoya. Political disapproval of his involvement in a socialist study group led to his leaving Japan for the United States in 1913.

After receiving a Ph.D. at Harvard in 1940, he taught there briefly as a lecturer. During that period, he married Masako Wada, the niece of Marquis Koichi Kido. They left the United States for Japan in 1942 on board a ship exchanging citizens during wartime.

During the reconstruction of Japan after the Second World War, he served first in the Ministry of Foreign Affairs and later as the vice minister of the Economic Stabilization Board, where he took part in the preparation of the first issue of the Board's Economic White Paper.

In 1948, he was appointed Professor of Economics at Hitotsubashi University, where he later served for nine years as the director of the Institute of Economic Research and for three years as president of the university from 1972 until retirement. After retirement, he served as an adviser to the Asahi newspaper. He then assumed a professorship at Meijigakuin University.

Tsuru's analytical works in economics are based on his wide background in the area of both Marxian and modern economics. His principal studies have been incorporated in the *Collected Works of Shigeto Tsuru* (1976). These constitute 13 volumes, and the last one, in English, is entitled *Towards a New Political Economy.*

The main areas of the author's interests which developed from his Harvard days encompassed Marxian methodology, business cycle theories and their application to Japan's economic development. A continuing emphasis in the studies was on aspects of the development of capitalism in Japan. His book *Has Capitalism Changed*? (1961) reflects this particular interest.

Tsuru was one of the first economists in Japan to have drawn the attention of the general public to environmental problems by applying his unusual skills in putting academic concepts into the language of ordinary people.

Selected Works

1961. *Has capitalism changed*? Tokyo: Iwanami Shoten Publishers.
1976. *Collected works of Shigeto Tsuru,* 13 vols. Tokyo: Kodansha. Vol. 13*, Towards a new political economy*, is in English.

Tucker, George (1775–1861)

Henry W. Spiegel

American economist and statistician, he was born in Bermuda, the offspring of a family prominent there and in Virginia. When he joined the faculty of the University of Virginia in 1825, he had already made a name for himself as a lawyer, man of letters and member of Congress.

Among Tucker's economic writings, *The Laws of Wages, Profits and Rent, Investigated* (1837) stands out as a theoretical contribution. Instead of Ricardo's labour theory of value he proposes a supply and demand theory. Ricardo had related rent to the original and indestructible powers of the soil, while according to Tucker rent arises because land yields a surplus. Ricardo had taught that profits tend to decline with rising money wages. To Tucker, the decline of profits reflected the movement from high-yielding investment projects to low-yielding ones. Wages, according to Ricardo, would remain at subsistence level. Tucker did not deny this, but called attention to the diminished quality of subsistence, which, as population grows, shifts to nutrients that require less land; for example, from meat to cereals to potatoes.

Tucker's pioneering work in statistics and demography, *Progress of the United States in Population and Wealth in Fifty Years as Exhibited by the Decennial Census* (1843), includes an early estimate of the national income – $62 per capita – and the prediction of the 'euthanasia' of slavery once the institution would no longer pay for itself.

Tucker's contributions demonstrate the high achievements of the Southern contingent among *antebellum* economic writers in the United States. His views about slavery, free trade and manufacture diverged from the pattern characteristic of the South and attest to the independence of his mind.

Selected Works

1837. *The laws of wages, profits and rent, investigated*. Philadelphia: E.L. Carey & A. Hart.
1843. *Progress of the United States in population and wealth in fifty years as exhibited by the decennial census*. New York/Boston: Press of Hunt's Merchant's Magazine/Little & Brown.

References

Dorfman, J. 1946. *The economic mind in American civilization 1776–1865*, vol. II. New York: Viking.
Dorfman, J. 1964. George Tucker and economic growth. In *The theory of money and banks investigated,* ed. G. Tucker, first published 1839, reprinted. Clifton: Kelley.
Snavely, T.R. 1964. *George Tucker as political economist*. Charlottesville: University Press of Virginia.

Tucker, Josiah (1713–1799)

G. Shelton

Keywords
Free trade; Hume, D.; Self-interest; Tucker, J.

JEL Classifications
B31

Born in Laugharne, Carmarthenshire, Tucker was Dean of Gloucester from 1758 until his death, and was also a rector in Bristol for over 50 years. Although his career as an ecclesiastic was a long and honourable one, he was best known in his

own day for his active part in many contemporary controversies. Whether the subject was the naturalization of foreign Protestants and Jews, the undesirable effect of low-priced liquors, or the cruel custom of cock-throwing on Shrove Tuesday, his pen was always ready. He was responsible for the earliest study of the Methodist movement and the first substantial critique of Locke's political philosophy. The themes which recurred most often were his opposition to monopolies and his hatred of war. His interest in political affairs was not confined to the press: he participated in several Bristol elections as the local Whig agent.

Tucker's period of greatest notoriety came during the American Revolution. In a steady stream of publications he rejected both the conciliation policy of Burke and that of war. Although he had no sympathy for the ideas espoused by the more radical Americans and their supporters in Britain, he saw no economic reason for attempting to retain the colonies by force, since he was convinced that they would willingly trade with her as long as it was in their interest to do so.

In his *Essay on Trade* (1749) Tucker recognized the need for a scientific study of what is now called economics but only the first part of what to be his 'great work' on the subject was ever printed, and then only for circulation among friends. However, his other works, which contained the bulk of his ideas, were known to Quesnay and Turgot (who translated one of them) well before the Physiocrats' first writings appeared, and several of his books were to be found in Adam Smith's library. These ideas included: self-love as a socially useful drive, labour as the true source of wealth, and the importance of machinery as a means of increasing that wealth. His aim was to encourage high productivity which would lead, in turn, to lower prices, increased demand, and more jobs. Anything which obstructed the free circulation of labour and capital, especially regulations supporting vested interests, should be eliminated. On the other hand, Tucker did not expect that self-interest and the public good would always coincide; some legislation was necessary to encourage that happy outcome by making what was socially desirable also profitable.

Tucker's most significant contribution may have been his argument against the notion put forward by David Hume that rich countries were likely over time to lose their wealth to poorer ones. Tucker eventually convinced Hume that the factors which made a nation rich in the first place tended to give it a practically insurmountable advantage over its less wealthy neighbours. Since Britain enjoyed such an advantage, once Tucker's reasoning was accepted, as it was by Pitt in the 1780s, opposition to free trade could be disarmed and the way cleared for its triumph in the 19th century.

Selected Works

1749. *A brief essay on the advantages and disadvantages which respectively attend France and Great Britain with regard to trade [The essay on trade]*.
1751–2. *Reflections on the expediency of a law for the naturalisation of foreign Protestants*. Part I (1751), Part II (1752).
1753. *Letters to a friend concerning naturalisation.*
1755. *The elements of commerce and theory of taxes*. In Schuyler (1931).
1757. *Instructions for travellers*. In Schuyler (1931).
1774. *Four tracts together with two sermons on political and commercial subjects*. Gloucester/London: Raikes & Rivington.
1775. *A letter to Edmund Burke*.
1781. *A treatise concerning civil government*. London: T. Cadell.

Bibliography

Clark, W.E. 1903. *Josiah Tucker, economist*. New York: Columbia University Press.
Schuyler, R.L. (ed.). 1931. *Josiah Tucker: A selection from his economic and political writings*. New York: Columbia University Press.
Semmel, B. 1965. The Hume-Tucker debate and Pitt's trade proposals. *Economic Journal* 75: 759–790.
Shelton, G. 1981. *Dean Tucker and eighteenth-century economic and political thought*. London: Macmillan.

Tugan-Baranovsky, Mikhail Ivanovich (1865–1919)

Alec Nove

Keywords

Exploitation; Labour theory of value; Marxist political economy; Serfdom; Subjective theory of value; Surplus value; Tugan-Baranovsky, M. I.; Underconsumptionism

JEL Classifications

B31

Of mixed Ukrainian-Tartar origin, Tugan-Baranovsky was born in the Kharkov province, and graduated from Kharkov university in 1888. His *Magister* dissertation for Moscow University was on industrial cycles in Great Britain, and he spent six months of his research time in London in 1892. There could scarcely have been a more masterly master's thesis. It was published in 1894. While criticizing crude underconsumptionist theories, and pointing out that 'the process of production creates its own market', especially for producers' goods, he went on to stress that the simple model derived from J.-B. Say assumes that 'that entrepreneur, before beginning production, has a wholly correct and accurate knowledge of the requirements of the market and of the output of every branch of industry'. He cited Moffat's phrase 'the continuous struggle between the requirements of unknown demand and the fluctuations of unknown supply'. He contrasted the 'propensity to save' with the output of capital goods of various types, and with the opportunities to invest, which can and do get out of line with one another. He collected much empirical data. In the words of Alvin Hansen, 'he began a new way of thinking about the problem' of business cycles.

His doctoral dissertation was another masterpiece, full of original research, *The Russian Factory, Past and Present (Russkaya fabrika v proshlom i nastoyashchem)* (1898), which has appeared in English translation. This was a major contribution to economic history. In vivid and well-written pages, Tugan-Baranovsky shows the great importance of the state and of serfdom, and the subsequent growth of market-orientated industries based on free labour (though some workers were serfs on quitrent, a few of whom became serf millionaires). He also made stimulating observations concerning 'natural' and 'artificial' industrialization, relevant to today's concerns with economic development.

His major contribution to economic theory was *Osnovy politicheskoi ekonomii* (1917), which went through many editions, and represented an attempt at a synthesis between Marxist political economy (the labour theory of value) and subjective value theory. He considered that the marginalists ignored 'the objective conditions of production', while Marxists failed to recognize that not only objective factors but also subjective valuations were an integral part of a theory of value. He argued that Marx confused value (*Wert*) with cost (*Kosten*). He basically supported Marx's theory of exploitation, but defined 'surplus value' as equal to the value of the *products* acquired (consumed) by the capitalists, which earned him criticism from Kondratiev and Struve. He retained from his early Marxism the belief that economists should regard man as not just another factor of production. If horses could write economics, there would be a horse theory of value.

Tugan-Baranovsky to the end of his life retained a particular interest in agriculture and in (voluntary) cooperation. One of his last articles drew attention, prophetically, to the effect of the egalitarian land redistribution of 1917–18 on the marketing of foodstuffs.

His academic career was mainly in the University of St Petersburg, though he was dismissed in 1899 for 'political unreliability' and only reinstated in 1905, as a *privatdozent*. His election to the chair of political economy in 1913 was vetoed by the Minister of Education. Re-elected in 1917, he did not take up his appointment, but returned to his native Ukraine. He became Academician, dean of the Faculty of Law of Kiev, chairman of the Ukrainian cooperatives, president of the Ukrainian economic association, and for a short

period Minister of Finance, amid turmoil and civil war. He died in 1919, on his way to Odessa to board a ship for France. He must be seen as the most original of the Russian economists of his generation. Alas, in the Soviet Union he was known chiefly as a 'legal-marxist' opponent of Lenin, and few had the opportunity to study his works, though *The Russian Factory* was reprinted.

Selected Works

1894. *Promyshlennye Krizisy v sovremennoi Anglii* [Industrial crises in contemporary Britain]. St. Petersburg. 2nd Russian ed. Trans. into French by J. Schapiro as *Les crises industrielles en Angleterre*, Paris: M. Giard & E. Briere, 1913.

1898. *Russkaia fabrika v proshlom i nastoiashchem* [The Russian factory, past and present]. St. Petersburg. 3rd Russian ed. Trans. A. Levin and C.S. Levin, under the supervision of G. Grossman, as *The Russian Factory*, Homewood, IL: R.D. Irwin, for the American Economic Association, 1970.

1905. *Teoreticheskie osnovy marksizma* [The theoretical foundations of Marxism]. St. Petersburg. Trans. into German as *Theoretische Grundlagen der Marxismus*, Leipzig: Duncker & Humblot, 1905.

1906. *Souremennyi sotsializm v svoem istoricheskom razuitii*. Trans. M.I. Redmount as *Modern Socialism in its Historical Development*, London: S. Sonnenschein & Co., 1910. Reprinted New York: Russell & Russell, 1966.

1914a. *Ekonomicheskaia priroda kooperativov i ikh klassifikatsiia* [The economic nature of cooperatives and their classification]. Moscow.

1914b. *Ocherki iz noveishei istorii politicheskoi ekonomii i sotsializma* [Outlines of the recent history of political economy and of socialism]. St. Petersburg.

1917. *Osnovy politicheskoi ekonomii* [Foundations of political economy]. Petrograd.

1918. *Sotsializm kak polozhitelnoe uchenie* [Socialism as a positive subject]. Petrograd.

Tugwell, Rexford Guy (1891–1979)

Leon H. Keyserling

Rexford Guy Tugwell was born in 1891. As an undergraduate at the Wharton School, he was jarred by textbooks reciting classical economic theories but ignoring real life; he wrote:

> I am sick of a nation's stenches,
> I am sick of propertied czars,
> I will roll up my sleeves,
> – make America over!

This did not bespeak revolutionary inclinations; it merely noted the unacceptable and his resolve to do something about it.

As head of the Economics Department in Columbia College, the undergraduate school in Columbia University, Tugwell stood against the graduate economics faculty; he insisted that more of the best economists should be assigned to teach undergraduates who later would vote upon or directly make public policies, rather than being used excessively in the graduate faculty to shape economics professors. And to counteract classical texts, he wrote for classroom use an institutional study, *American Economic Life and the Means of Its Improvement*. And, *The Industrial Discipline and the Governmental Arts*, showing preference for national economic planning.

During F.D.R.'s last two years as Governor of New York, Tugwell joined in advising how a President should fight the Great Depression. His revealing book, *The Brains Trust,* depicted F.D.R.'s innate conservatism but also made clear that his advisers prompted him toward boldness and experimentation.

Tugwell next served as Assistant and then Undersecretary in the Department of Agriculture. In these posts, his fight for improved pure food and drug laws stirred up violent opposition, and his outspoken liberalism made him a whipping boy for the Administration. More proximately, his siding with Jerome Frank, General Counsel of the Agricultural Adjustment Administration, and

others on behalf of small rather than the large farmers championed on Capitol Hill led to his being transferred to head the Resettlement Administration, initiating towns like Greenbelt, Maryland.

However, his most vital role in Washington was as leader of an informal but influential group favouring centralized national planning, rather effective during the 'first New Deal'. A second informal group, led intellectually by Louis D. Brandeis (and fortified by his Supreme Court votes) and Felix Frankfurter, turned action considerably toward the 'second New Deal', marked by a slowdown of strong policies, and with much inveighing against 'the curse of bigness' and 'economic royalists'.

Rex's ideas were in no way totalitarian; they urged peacetime application, with appropriate modifications, of the type of comprehensiveness economic planning later used during World War II, with large interpenetration between Government and business. Sensing increasing frustration and official rejection, Tugwell left the Government to become Vice-President of American Molasses Company and the Chairman of the New York City Planning Commission. Roosevelt then appointed him the first Governor of Puerto Rico, with good accomplishments as reported in *The Stricken Land*.

Thereafter, Tugwell taught planning at the University of Chicago, and finally spent many years at the Hutchins Center for the Study for Democratic Institutions in Santa Barbara. His views never altered, that New Deal 'patchwork' would never work and that centralized national planning was imperative. Coming to feel that current trends in economics were hopeless and incorrigible, he turned to institutional proposals for rearranging the structure of government. Hence, his monumental work, *The Emerging Constitution*. He knew that none of his recommendations would be adopted, but felt that stating them would be of value. His many books about the Presidency included *The Democratic Roosevelt,* beautifully written like all his work, and an objective and critical evaluation of one whom he admired endlessly. It is probably the most revealing book about F.D.R.

At memorial services in 1979, I spoke of a man who, among all whom I got to know during a half century of public service, was truly one of 'the best and the brightest'.

Selected Works

Tugwell's publications are far too numerous to be listed here, but those listed are representative of his work and interests.

1922. *The economic basis of public interest.* Menasha: George Banta Publishing Company.
1924. *The trend of economics* (editor and contributor). New York: Knopf.
1925. (With Thomas Monroe and Roy E. Stryker.) *American economic life and the means of its improvement.* New York: Harcourt, Brace.
1927. *Industry's coming of age.* New York: Harcourt, Brace.
1932. *Mr. Hoover's economic policy.* New York: John Day.
1933. *The industrial discipline and the governmental arts*. New York: Columbia University Press.
1934. (With Howard C. Hill.) *Our economic society and its problems.* New York: Harcourt, Brace.
1934–5. *Redirecting education* (ed. with Leon H. Keyserling, and contributor). New York: Columbia University Press. Vol. 1, 1934; Vol. 2, 1935.
1947. *The Stricken Land: The story of Puerto Rico*. Garden City, New York: Doubleday.
1957. *The Democratic Roosevelt.* Garden City, New York: Doubleday.
1967. *The light of other days.* Garden City, New York: Doubleday.
1967. *F.D.R.: Architect of an era.* New York: Macmillan.
1968. *The brains trust.* New York: Viking.
1974. *The emerging constitution.* New York: Harper's Magazine Press.
1982. (Posthumous.) *To the lesser heights of Morningside.* Philadelphia: University of Pennsylvania Press (Introduction by Leon H. Keyserling).

Tulipmania

Peter Garber

Abstract

This curious speculation in tulip bulbs near four centuries ago has become a modern synonym for and warning about the irrational speculation that may break out in any asset market. Nevertheless, a look at the forces that actually drove it provides an alternative, fundamental explanation and a warning to observers of asset markets not to accept so quickly unprovable psychological explanations of asset prices.

Keywords

Asset pricing; Bubbles; East India Company; Forward contracts; Multiple equilibria; Mississippi Bubble; South Sea Bubble; Speculative bubbles; Tulipmania

JEL Classifications

E3

The Netherlands of 1634–7 was the scene of a curious speculation in tulip bulbs that has come to be known as the Dutch tulipmania. Single bulbs of rare and prized varieties such as Semper Augustus or Viceroy became worth a middle-sized fortune. In its most extreme final phase in January-February 1637, prices of even common varieties such as Switsers or Witte Kroone soared twentyfold within a month and then crashed back to their original values. That these were prices of easily reproducible horticultural products has added to the bemusement of generations of historians and economists.

In the succeeding 370 years, the *historical* tulipmania became, in itself, an obscure footnote to the *conceptual* tulipmania of economics and finance, a word warning of the obvious, delusional speculative excess that human behaviour in financial markets can create (see, for example, Kindleberger 1996). It is interchangeable with words like 'bubble' or 'mania', which also arose from historically distant events such as the Mississippi or South Sea Bubbles or the more recent 'irrational exuberance'. These words have been used by economic theorists to emphasize an historical basis for the salience of unstable multiple equilibria in forward-looking financial and macroeconomic theories. They have also been used to justify ignoring financial market outcomes that contradict favoured asset pricing theories by means of an arbitrary invocation of the existence of a bubble.

The Traditional Image of Tulipmania

Modern references to the tulipmania usually depend on the brief description in Charles Mackay's *Extraordinary Popular Delusions and the Madness of Crowds* (1852). The tulip originated in Turkey and spread into western Europe in the mid- 16th century. The tulip was immediately accepted by the wealthy as a beautiful and rare flower, appropriate for the most stylish gardens. The market was for durable bulbs, not flowers. The Dutch dominated the market for tulips, initiating the development of methods to create new flower varieties. The bulbs that commanded high prices produced unique, beautifully patterned flowers; common tulips were sold at much lower prices.

Beginning in 1634, non-professionals entered the tulip trade in large numbers. According to Mackay, individual bulb prices reached astronomic levels. For example, a single Semper Augustus bulb was sold at the height of the speculation for 5,500 guilders, a weight of gold equal to $66,000 evaluated at $600/oz. Mackay provided neither the sources of these bulb prices nor the dates on which they were observed, however.

Finally, and unexplained by Mackay, the frenzy suddenly terminated. According to Mackay, even rare bulbs could find no buyers at ten per cent of their previous prices, creating long-term economic distress. Mackay presented no evidence of immediate post-collapse transaction prices of the rare bulbs. Instead, he cited prices from bulb sales of 60 years, 130 years, or 200 years later as indicators

of the magnitude of the collapse and of the obvious misalignment of prices at the peak of the speculation. Moreover, Mackay provided no evidence of the general economic context from which the speculation emerged.

The Fundamentals of the Tulipmania

Unfortunately, the fundamentals of markets in rare bulbs present a much more prosaic picture. The bulk of the speculation concerned highly prized tulips that were infected with mosaic virus. Mosaic virus had the effect of producing unique feathery patterns in the flower that could be reproduced only through propagation by budding, not by seeds. Hence, the rate of reproduction was much more limited that one might expect. Such bulbs were traded primarily among professionals. Their prices were supported by a strong demand by flower fanciers, not only in the rapidly growing Netherlands of the golden age but also by the wealthy nobility and merchants of surrounding countries. During the period of the tulipmania, 1634–7, the already high prices of such bulbs doubled or tripled. Over the course of decades or centuries, prices for these varieties converged to the low cost of reproduction, and this has been taken as evidence of folly.

However, an examination of the pricing of prized flower varieties throughout history reveals a similar pattern: prices of the prototype are very high, perhaps even representing a medium fortune. Then, as they are reproduced through succeeding generations, they become common. Just as for the value of a prized racehorse put out to stud, the high initial price represents the present discounted value of the valuation above cost of successive, expanding generations, wherein the value of any individual exemplar is bound to fall.

The more frenzied phase of the tulipmania described by Mackay took place from mid-1636 to February 1637, but especially in January, 1637. At this time trading, especially among the non-professionals, took place in newly organized 'colleges', which were located in taverns. The trading was not for actual bulbs but for contracts for forward delivery. Since bulbs had to remain in the ground through the winter, none were actually delivered on these contracts before the speculation ended. Contracts were not marked to market, and margin was not posted. A small, fixed amount of 'wine money' had to be delivered by the buyer, which provides the flavour of, if not the fuel for, what was happening during the frenzied trading in these taverns.

When this part of the speculation collapsed in February 1637, some city governments proposed winding up outstanding contracts with a ten per cent payment on contracted amounts if the buyer refused to accept delivery. Perhaps this is where Mackay got the notion that bulbs could not be sold at ten per cent of their previous value, even though a buyer might refuse the deal if prices had fallen only to 90 per cent of the contracted amount. There were very few takers even on this offer, but short sales were in any case unenforceable contracts under Dutch law.

When one looks at notarized contracts for actual bulbs, however, the picture is quite different. Some rare bulbs that were auctioned for high prices at the very peak of the speculation in February 1637 still sold for high, albeit much lower, prices in 1642. For example, an Admirael Liefkens bulb was sold for 1,345 guilders at the peak and for 220 guilders in 1642, an annual percentage decline in value of 36 per cent. This rate of decline is comparable to the typical pattern of price behaviour for valued varieties in successive historical periods and does not indicate anything unusual in the mania.

It is the rare bulb price behaviour during the tulipmania that has been emphasized historically. But at the very end common bulbs sold in bulk shot up twentyfold and soon collapsed back to one-twentieth of the peak. It is this usually ignored bit of the episode that remains a puzzle.

An Historical Background

The tulip market was introduced into the Netherlands during the Eighty Years' War of independence between the Dutch and the Spanish, and the tulipmania occurred in the middle of the Thirty Years' War as the two conflicts merged.

The Spanish were thwarted in their attempts to subjugate the Netherlands, which consolidated its territory and eventually seized control of most of international shipping. The Thirty Years' War of 1618–48 was particularly destructive of the populations and economies of central Europe, with many principalities in the Holy Roman Empire losing one-third of their populations.

In every year of the war, the Dutch fielded large armies and supported large fleets, though the population of the Netherlands was no more than 1.5 million. The Dutch provided much of the strategic planning and finance for the Protestant effort, along with France, negotiating and financing the successive interventions of Denmark and Sweden on the Protestant side in the 1620s and 1630s.

From 1620 to 1645, the Dutch established near-monopolies on European trade with the East Indies and Japan, conquered most of Brazil, took possession of the Dutch Caribbean islands, and founded New York. In 1635 the Dutch formed a military alliance with Richelieu's France, which eventually placed the Spanish Netherlands in a precarious position. In 1639 the Dutch completely destroyed a second Spanish Armada of a size comparable to that of 1588. As a result of the war, Spain ceased to be the dominant power in Europe, and the Netherlands, though small in population and resources, temporarily became a major power centre because of its complete control over international trade and international finance. The Dutch were to 17th-century trade and finance as the British were to 19th-century trade and finance.

Sophisticated finance mechanisms evolved with the establishment of its trade and finance dominance. Amsterdam became the leading market for short- and longterm credit; and markets in stocks, commodity futures, and options materialized early in the 17th century. Trading of national loans of many countries centred on Amsterdam, as did a market in the shares of joint stock companies. The East India Company, founded in 1602, gradually gained control over east Asian trade and consistently paid out large dividends. Interest rates on Dutch markets were remarkably low for the times; for example, the East India Company paid no more than five per cent on advances during the 17th century.

There were some dark periods during this golden age, and it should be carefully noted that these occurred during the years of the tulipmania. From 1635 to 1637, bubonic plague ravaged the Netherlands. In July 1634 the Holy Roman Empire completely defeated Swedish forces in the Battle of Nordlingen, forcing a treaty on the German Protestant principalities in the May 1635 Peace of Prague and releasing Spanish resources for the war against the Dutch. Along with the growing war weariness in the Netherlands, these events forced France to enter the Thirty Years' War militarily with the Dutch alliance in 1635. Initially unprepared, the French suffered major setbacks, culminating in an imperial invasion of northern France in August 1636.

How Should We Interpret the Tulipmania?

The tulipmania is an obscure event from distant history that provides a cornucopia of concepts even now. One can take one's pick from the following views, depending on personal taste:

- It was an outburst of speculative fever that serves to the present day as a warning of the dangers of market speculation.
- It was a curious event limited to the Dutch winter of 1636–7 in the middle of an outbreak of bubonic plague and at the time of the greatest success of the Catholic armies of the Empire in the Thirty Years War against the Protestants.
- It was a drinking game in which people without wealth made the equivalent of million euro bets with each other, with no intention or possibility of paying.
- It was a swing of fashion in the most wealthy society of the era, which caused the most exquisite of tulips to have a higher price than Rembrandt's *Night Watch.*
- It was a reasonable and well-calculated investment that still causes the most wonderful outburst of colour every Dutch springtime.

See Also

► Speculative Bubbles

Bibliography

Cooper, P. 1970. *New Cambridge modern history, volume 4: The decline of Spain and the Thirty Years' War*, 1970. Cambridge: Cambridge University Press.
Garber, P. 2000. *First bubbles: The fundamentals of early manias*. Cambridge, MA: MIT Press.
Kindleberger, C. 1996. *Manias, panics and crashes: A history of financial crises*. 3rd ed. New York: Wiley.
Mackay, C. 1852. *Extraordinary popular delusions and the madness of crowds*. Vol. 2. 2nd ed. London: Office of the National Illustrated Library.
Posthumus, N. 1929. The tulip mania in Holland in the years 1636 and 1637. *Journal of Economic and Business History* 1: 434–466.
Schama, S. 1987. *The embarrassment of riches*. New York: Alfred Knopf.

Tunisia, Economy of

Barry Turner

Keywords
Jasmine revolution; Millemes; Multi-Fibre Agreement; Tunisian dinar

JEL Classification
O53; R11

Overview

Tunisia's economic record compares favourably with other developing countries, particularly those on the African continent. After a balance of payments crisis in the mid-1980s, steps were taken to improve macroeconomic policy, foster the non-public sector and liberalize prices and controls. In the 1990s the economy grew steadily at an average annual rate of 5%. Inflation and annual budget deficits have fallen and poverty has been reduced.

Tunisia's trade is oriented towards the EU, with France, Italy, Germany and Spain the country's most important trading partners in descending order. However, Tunisia's textile exporters lost EU market share to Asia as a result of the end of the Multi-Fibre Agreement quota system in January 2005. The economy is diversified relative to many of its non-European neighbours, with significant mining, tourism, energy, manufacturing and agricultural sectors.

In 2008 manufacturing contributed 18.5% to GDP; followed by trade and hotels, 14.5%; finance and real estate, 12.2%; transport and communications, 11.2%; and public administration and defence, 10.5%.

A swift policy response left the economy largely unaffected by the global financial crisis. Growth was 3.1% in 2009, with exports and domestic demand rebounding. Following the Jasmine Revolution of early 2011 and the neighbouring Libyan crisis, the short-term economic outlook weakened, with tourism and foreign direct investment suffering.

Unemployment remains persistently high, particularly among young people, with the unemployment rate for university graduates at 25%. In May 2011 the G8 group of countries promised US$20 bn. in loans and grants to Tunisia and Egypt over a three-year period. In 2011 the interim government announced a US$1.5 bn. stimulus package, plus an emergency economic and social development plan. This includes measures to tackle security problems and unemployment, boost private sector growth and regional development and improve the lot of the poorest families.

Currency

The unit of currency is the *Tunisian dinar* (TND) of 1,000 *millimes*. The currency was made convertible on 6 January 1993. Foreign exchange reserves were US$4,069 m. and gold reserves 218,000 troy oz in July 2005. Inflation was 4.4% in 2010 and 3.5% in 2011. Total money supply was 8,339 m. dinars in June 2005.

Budget

The fiscal year is the calendar year. Budgetary central government revenue totalled 13,266 m. dinars in 2008 and expenditure 11,544 m. dinars. Taxes accounted for 85.4% of total revenues in 2008. Principal sources of revenue in 2008: taxes on goods and services, 5,061 m. dinars; taxes on income, profits and capital gains, 4,561 m. dinars; taxes on international trade and transactions, 965 m. dinars. Main items of expenditure by economic type in 2008: compensation of employees, 5,164 m. dinars; subsidies, 2,713 m. dinars; interest, 1,143 m. dinars.

VAT is 18% (reduced rates, 12% and 6%).

Performance

Real GDP growth was 3.1% in both 2009 and 2010 but the economy contracted by 1.8% in 2011. Tunisia's total GDP in 2012 was US$45.7 bn.

Banking and Finance

The Central Bank of Tunisia (*Governor*, Chedly Ayari) is the bank of issue. In 2003 there were 12 commercial banks, six development banks, two merchant banks and five 'offshore' banks.

In 2010 external debt totalled US$21,584 m., equivalent to 51.1% of GNI.

There is a small stock exchange (51 companies trading in 2007).

See Also

- ▶ Energy Economics
- ▶ International Monetary Fund
- ▶ Islamic Economic Institutions
- ▶ Islamic Finance
- ▶ Oil and the Macroeconomy
- ▶ Organization of the Petroleum Exporting Countries (OPEC)

Turgot, Anne Robert Jacques, Baron de L'Aulne (1727–1781)

Peter Groenewegen

Keywords

Advances; Böhm-Bawerk, E. von; Capital accumulation; Capital and interest theory; Class; Condorcet, Marquis de; Division of labour; Du Pont de Nemours, P. S.; Gournay, Marquis de; Inequality; Interest; Laissez-faire; Law of variable proportions; Loanable funds; Marginal revolution; Market price; Natural price; Physiocracy; Quesnay, F.; Saving Equals Investment; Sharecropping; Single Tax; Slavery; Smith, A.; Stages theory of progress; Surplus; Thrift; Turgot, A. R. J.

JEL Classifications

B31

Economist, philosopher and administrator, Turgot was born in Paris, in 1727, the third son of a well-established Norman family with a long tradition of public service in the magistrature. Destined originally for a career in the Church, his education was extensive. Because of shyness, his education commenced at home with a private tutor; it continued at the Collèges Duplessis and Bourgogne where, among other things, he studied the philosophical systems of Newton and Locke. In October 1746 he entered the Seminary of Saint-Sulpice in preparation for the priesthood. From June 1749 to early 1751 he was resident student at the Maison de Sorbonne, an annex of the theological faculty of the University of Paris. His already considerable academic distinction led to his election to the office of prior in 1750. This honorary position inspired two of his earliest works, of which the second, *Philosophical Review of the Successive Advances of the Human Mind* (Turgot 1750a) contained a demonstration of the importance of economic surplus for the development of

civilization as part of his four stages theory of human progress:

> Tillage ... is able to feed more men than are employed in it ... Hence towns, trade, the useful arts and accomplishments, the division of occupations, the differences in education, and the increased inequality in the conditions of life. Hence leisure... (and) the cultivation of the arts. (Turgot 1750a, p. 43)

His father's death in early 1751 possibly saved Turgot from having to take his final vows, since his inheritance provided sufficient income to commence the administrative career he desired. He gained appointment to some judicial positions, including that of Master of Requests in early 1753, the stepping stone to a career as provincial intendant. During the 1750s, Turgot's prolonged residence in Paris allowed immense intellectual activity but left time for extensive travels through France when accompanying Gournay on his official tours of inspection of French industry. His contributions to the *Encyclopédie* (Turgot 1756; 1757a; 1757b), ranging from articles on Etymology, Existence and Expansibility to Fairs and Foundations, spread his fame as a philosopher and gained him the friendship of Voltaire, whom he visited in 1760 during his only journey abroad.

Although Turgot's interest in economics had commenced as early as 1749 when he wrote a critique of Law's system of paper currency, and such interest was maintained in the Sorbonne orations and other early writings, it was considerably stimulated by Gournay's friendship. Gournay's death inspired Turgot's famous eulogy (Turgot 1759) and earlier he had encouraged Turgot's translation of Tucker (Turgot 1755), Turgot's comments on Gournay's notes to the translations of Child (Turgot 1753–4), and most probably, aspects of the content of Turgot's two economic articles for the *Encylopédie* (1757a; 1757b). Gournay's friendship was particularly important because it brought Turgot's economics under more substantial English influence as compared with Physiocracy (Groenewegen 1977, p. xiv). Turgot's first meeting with Quesnay cannot be precisely dated. It may not have occurred till 1756 or 1757 when their mutual association with the *Encyclopédie* may have brought them together. Turgot (1759, p. 26) cites Quesnay's contributions with considerable approval, indicating that his generally good relations with the Physiocrats must by then have been well established. His presence at Quesnay's meeting in the Entresol at Versailles as a 'handsome young Master of Requests' was in any case recorded by Mme de Hausset (n.d., pp. 117–19). A lifelong friendship with Du Pont de Nemours, which began in 1763, must have strengthened good relations with the Physiocrats even further. Another enduring friendship was made with the philosopher and mathematician, Condorcet. Both friends produced memoirs of Turgot's life after his death (Du Pont 1782; Condorcet 1786).

In 1761 Turgot was appointed Intendant of Limoges, a large district containing most of the provinces of Limousin and Angoumois, and this position he filled with distinction for 13 years. The task of the 18th-century intendant, a post compared by Morley (1886, p. 112) to that of Chief Commissioner in a large district of the former British Empire in India, were many:

> He had to collect direct taxation, rectify justice, promote the arts of agriculture, encourage industry and commerce ... Everything came within the scope – sanitation and public order, morality and poor relief, the recruiting and billeting of soldiers, military equipment, rations and transport, religious processions and the pairs of churches, colleges and libraries, parochial and municipal finance. (Dakin 1939, pp. 27–8, the standard source for details of Turgot's administrative career as intendant)

Despite this cumbersome and heavy administrative load, Turgot managed to introduce some reforms. These included changes to the assessment and collection of the taille, transmutation of the *corvée* and the *milice* into money payments, and establishing public workshops to alleviate hardships suffered by the population of his province during the long and severe famine of 1769–1772. Many of his better known economic writings date from this period: first of all, his *Reflections on the Production and Distribution of Wealth* (Turgot 1766a, but not published till 1769–70 in serial form in the *Ephémérides*), a draft for a paper on value and money (Turgot 1769), observations on two winning entries in a prize competition he had organized on the subject

of taxation (Turgot 1767a; 1767b), and a series of memoranda connected with the ministration of his province in which he pleaded with the central authorities for reforms on the basis of carefully elucidated theoretical principles. The more important of these deal with taxation in general (Turgot, 1763), mines and quarries (Turgot, 1767a), the grain trade (Turgot, 1770a), the rate of interest (Turgot, 1770b) and the trade mark on iron products (Turgot, 1773). His administrative work permitted regular but infrequent visits to Paris to see friends and attend the salons of Mme de Graffigny, Mme de Geoffrin and later, Mlle de Lespinasse. Apart from the French intelligentsia, he there became acquainted with foreign notables like Hume, Adam Smith, Franklin and Gibbon. The exile imposed by his administrative position also inspired a substantial correspondence with Du Pont de Nemours, Condorcet, and his personal secretary, Caillard, making up a large part of the five volume edition of his works as edited by Schelle (1913–23). Schumpeter (1954, p. 248) notes a further significant aspect about Turgot's administrative career: 'nearly all his creative work must have been done between 18 and 34 because during the 13 years at Limoges, Turgot can have had but scanty leisure, during his nearly 2 years of ministerial office, practically none.'

Louis XVI's succession to the throne in 1774 marks the next stage in Turgot's career; his membership of the Royal Council, first as Minister of the Navy (from 20 July 1774), then as Minister of Finance (from 24 August 1774 to 12 May 1776, the date of his dismissal). While lamenting the fact that so much more could have been done, Du Pont (1782) summarized Turgot's career as minister in terms of the reforms accomplished. These included restoration of the domestic free trade in grain, abolition of many small, local duties and other constraints on trade, and the January 1776 measures, of which partial suppression of the guilds and replacing the *corvée* with a more general land tax were the more controversial measures. These last, now generally known as the six Edicts, ultimately caused his downfall even though he did secure royal support for their forcible registration at a famous *Lit de Justice*. As an 'experience in economic politics, an exception to the general rule that French ministers of finance are financiers rather than economists', Faure (1961) gives a detailed account of Turgot's ministerial experience and the opposition it encountered almost from the start during the grain riots of early 1775 and the intrigues surrounding the campaign to prevent registration of his 1776 Edicts. The reforms Turgot had accomplished were reversed within six months from his downfall, and 'leisure and complete freedom as the principal net product from my two years in the ministry' was how he himself sarcastically summed up his achievements in a letter to Caillard (Schelle 1913–23, vol. 5, p. 488). The period of retirement in the five years which remained of his life were not years of inactivity.

> The sciences which he had formerly cultivated, easily filled up his time; he studied mathematics, he sought to bring the thermometer to greater precision, he searched with l'Abbé Rochon, after various expeditious convenient, and cheap methods of multiplying copies of writing to supply the place of printing, . . . he preserved all his passion for literature and poetry . . . (Condorcet 1786, pp. 255–62)

In 1778 he was elected President of the Académie des Inscriptions et des Belles Lettres. He died in Paris in March 1781 from gout, a family illness that had steadily wrecked his health, worsening particularly during the last decade of life.

Although Turgot is now largely remembered as a very important 18th-century French economist and a pre-revolutionary reformist finance minister, such an assessment fails to reflect his youthful ambitions and work. Meek (1973, pp. 1–2) indicates that.

> Turgot set out from the beginning with the conscious intention of becoming a polymath rather than a specialist. . .. A list of works to be written . . . begins with 'The Barcimedes', a tragedy, and ends with 'On Luxury, Political Reflections', and in between these are forty-eight others, including works on universal history, the origin of languages, love and marriage, political geography, natural theology, morality and economics, as well as numerous translations from foreign languages, literary works, and treatises on scientific subjects . . . What is (especially) remarkable is . . . that Turgot managed, during his short life of only fifty-four years, to make some contribution to so many of them, or at least to retain an active and intelligent interest in them.

Turgot was therefore considerably more than an economist, and some of the qualities Meek listed needed to be highlighted to underscore that fact. In the first place, he was a superb linguist, 'reading seven languages – Greek, Latin, Hebrew, Spanish, German, Italian, and English, the last three of which he spoke fluently' (Dakin 1939, pp. 10–11) and from some of which he published poetry translations (Turgot 1760; 1762; 1778). This linguistic skill is reflected in his magnificent library, the catalogue of which (Tsuda 1974) demonstrates his ability to profit from the economics writings of other countries. Secondly, his wider interests influence the interpretation to be given to his economics. Turgot's contributions cannot be simply assessed in terms of his importance in fashioning certain parts of the marginalists' toolbox, as done, for example, by Schumpeter (1954). He is far more correctly depicted as 'an author of transition between the Physiocrats at the end of the eighteenth century and the English classical economists at the start of the nineteenth' (Bordes 1981, p. xvi), that is, the true contemporary of those like Smith, Steuart, Condillac, Verri and Beccaria producing economic treatises building on Physiocracy in that quarter century ending in 1776 during which political economy emerged as the science of the reproduction, circulation and distribution of wealth (see Groenewegen 1983a). Although, apart from the skeleton form of his *Reflections* (Turgot, 1766a), Turgot never completed such a treatise, this skeleton combined with his youthful views on social progress allows his economics to be depicted as something essential to the understanding of historical stages (see Finzi 1981). The reduced emphasis on the economics developed by him as a by-product of his administrative career this implies prevents his depiction as a 19th-century liberal (see Morley 1886; Bourrinet 1965) or as a general precursor of equilibrium theory (Nogaro 1944, pp. 26–7; Bordes 1981, pp. xxvi–xxviii).

Schelle (1913–23, vol. 1, pp. 29–30, 79) draws attention to the fact that the young Turgot was interested above all in sociology, shown by his attempts at analysing causes of progress and decay in taste, science and the arts, and that this analytical interest was enhanced by studying the formation of languages, because etymology provides valuable clues not only to the progression of ideas but to the needs from which ideas originate. Turgot's early work on language formation and social progress appears to have suggested the importance of economic factors in explaining this process, and that the means by which peoples gain their subsistence, determining as it does their access to economic surplus, is particularly important to explain the manner in which societies, morals, laws, the arts and the sciences gradually develop. Turgot (1750b, p. 172) explicitly relates certain characteristics in the formation of languages to stages of hunters, shepherds and husbandmen with their different requirements for communication. The notion of progress between these stages is implied in his critique (Turgot, 1751a, pp. 242–3) of the alleged virtues of equality in the savage state where he shows that, by preventing the division of labour and the accumulation of capital on which abundance and secure subsistence depend, such equalities also prevent progress in the sciences and arts. The subsequent fragment *On Universal History* (Turgot, 1751b) combines these elements into 'a quite advanced statement of the four stages theory – or at any rate a three stages theory, with a distinct hint of the fourth stage'(Meek 1977, p. 22).

Although Turgot's *On Universal History* was not completed, its basic notions stayed with him for the rest of his life and in a number of aspects received further development. The systematic attempt to explain general progress by stages from hunters, shepherds, farmers to a commercial society to a large extent provides the basis on which Turgot constructed his analysis of the production and distribution of wealth in the *Reflections*, as is explicitly recognized in his discussion of cattle as a form of moveable wealth (Turgot, 1766a, pp. 66–7). Moreover, the whole of the *Reflections* is imbued with Turgot's sociological concerns with nature of progress and historical development, thereby reinforcing the need to interpret its contents in terms of stadial development. Such a view is also appropriate for its alleged original purpose as providing explanations to accompany an extensive questionnaire on the Chinese economy and society which he

had prepared for two young Chinese students in 1766 (see Groenewegen 1977, pp. xvii–xix). This aspect of the *Reflections* may be demonstrated from a summary of its contents, a process facilitated by the parts into which Du Pont divided it for publication in *Ephémérides*.

Under this subdivision (Turgot, 1766a, pp. 43–56); the first part of the *Reflections* analyses the basic features of the production and distribution of wealth within an agricultural society. Although capital advances are used in such a society, the distributional aspect of such use is ignored at this stage except for its final sections dealing with what for Turgot were contemporary manifestations of agricultural production (1766a, pp. 56–6). Turgot argues at the outset that such an agricultural society presumes a division of labour, a natural consequence of its inequality of property ownership. Hence it presumes a specific set of class relations, that is, division of society between a proprietors' class owning land and living from its surplus produce without a need to work, and working classes without property earning their living from their labour. Within this working class, the division of labour divides those cultivating the soil to produce food and raw material or products of prime necessity from artisans who transform those primary materials into forms more suitable for people's use. Because artisans depend on those working in agriculture for their livelihood, Turgot calls them a stipendiary class. Because they only transform existing wealth without generating a surplus, Turgot calls them a sterile class to contrast their work with that in agriculture which produces such a surplus and thereby generates new wealth. In this way Turgot demonstrates the appropriateness of Physiocratic class analysis for understanding agricultural society.

At the end of this discussion Turgot suggests that the relationship between proprietors and the working classes in agriculture is itself subject to change with respect to the manner in which proprietors draw the surplus from the land through the organization of production. The method springing first to mind, landlords hiring wage labour for themselves, Turgot views as an unlikely candidate to be first from the perspective of actual historical development. Slavery appears to have been first in this regard. Though Turgot saw slavery persisting in colonial societies, elsewhere economic circumstances, combined with humanity and landlord's convenience gradually transformed slavery first into bondage of the soil and then vassalage, where former serfs become tenants and surplus product rent and other stipulated dues. One such tenancy, the dominant form in Turgot's France, was sharecropping or *métayage* in which the landlord made the advances in return for a fixed part of the produce; another, more advanced form existed where a capitalist farmer, or entrepreneur (Turgot, 1766b, pp. 28–9) rented the land from the proprietor for a specified rent and period of time, himself providing the necessary advances for cultivation. Capitalist farming or *la grande culture* had begun to emerge in France during Turgot's time, and the farmer/entrepreneur class it created 'has a quite distinct status from that of the ploughman/sharecropper. He does not earn his living by the sweat of his brow like labourers but by employing his capital in a lucrative manner like the shipowners of Nantes and Bordeaux employ theirs in maritime trade' (Turgot, 1766b, p. 29). As nations become more wealthy, and capital accumulates, a new proprietors' class is created who live without working from the revenue of money or capital. The section which opens the second part of the *Reflections* draws attention to this feature by dealing with 'capitals in general and the revenue of money'.

Explanations of the origin and use of capital, and its impact on 'the system of distribution of wealth which I have just outlined' (Turgot, 1766a, p. 56) requires an elementary acquaintance with the theory of money, commerce, exchange and value and hence some retracing of steps. After this digression, which contains little that is new, Turgot presents a fascinating analysis on both the uses of capital and its origins through accumulation and thrift. Turgot discusses accumulation and thrift both historically and analytically. Historically, accumulation is associated with slavery and surplus product from land: analytically, prudence and a desire for self-improvement are seen as major motives for thrift. Turgot argues that the savings process is greatly facilitated by the introduction of

money but that this raises new complications such as a need to distinguish saving, hoarding and investment. Turgot's saving–investment analysis denied the possibility that money savings were able to induce substantial leakages from the circular flow because hoarding was seen as irrational and money had only a limited role as a store of value. Turgot argued that savings were immediately transformed into investment (see Groenewegen 1971).

Turgot's analysis of the productive use of capital and its social implications is presented in the second and third parts of the *Reflections*. These reveal the degree to which his economics had departed from Physiocracy and anticipated views subsequently developed by Adam Smith. First of all, Turgot's exposition extends the use of capital to all sectors of industry thereby not confining it to agriculture as Quesnay had done. Secondly, Turgot, like Smith, links an increasing need for capital in production with extensions of the division of labour and a consequent lengthening of the time period of production. Thirdly, Turgot associates the provision of capital to industry with a new class of society, the capitalist/entrepreneur as owners of moveable wealth, who invest these resources to reap a return. Hence the working classes of agriculture, manufacturing and trade 'may be divided into two orders of men, that of the Entrepreneurs of Capitalists, who make all the advances, and that of the mere wage earning workmen' (Turgot, 1766a, pp. 72–3). Of special significance for analysing distribution, this new class appropriates the resources by which it can live without labour through the creation of interest and profit as a new income type. Profit in this context, is clearly associated by Turgot with a return on productive investment, comprising an interest component, a premium for risk and remuneration for the time and trouble of the entrepreneur in supervising the investment. Part of the *Reflections* therefore suggests that the quantitative changes of gradual capital accumulation (perhaps first experienced within agriculture) by a qualitative leap create a new stage of society, the commercial or capitalist stage (see Meek 1973, pp. 21–6).

However, this view is partially contradicted in some of the *Reflections*' later sections. These reveal the new class as mere lenders of money and show Turgot equivocating on whether interest and profit have the same disposable status as the net product of land. Such aspects of Turgot's work show that for him 'commercial society' perhaps remained 'incorporated into the agricultural state [never becoming] a separate stage, characterised and led by an internal logic of its own' (Finzi 1982, p. 116), and reinforce the position that Turgot's analytical schema is a transition from the Physiocrats to subsequent classical political economy retaining that ambiguous relationship between capital and land, rent and interest, not really resolved analytically until Ricardo's distribution analysis (see Cartelier 1981).

Despite this ambivalence in depicting the final stage of social progress, these sections of the *Reflections* also contain some of Turgot's most analytically significant contributions to economics. Having shown that 'capitals are the indispensable foundation of all lucrative enterprises' and that the continual reproduction of these capitals 'with a steady profit' constitutes 'the true idea of the circulation of money' the disturbance of which may cause economic decline (1766a, pp. 75–6), Turgot analyses the mutual interrelationship of the returns on various types of investment and the rate of interest. Interest itself is shown by Turgot to be determined by the demand for and supply of loanable funds, the demand arising from both consumption and investment needs. These investment needs, or employments of capital as Turgot calls them, are described as: purchasing a landed estate, which yields least; lending a capital at interest the return of which is greater; and investing in agricultural, manufacturing or commercial enterprises, the return of which is greatest. Irrespective of these inequalities in yield to the various employments of capital, Turgot argues that competition combined with capital mobility causes a tendency to equilibrium between them.

> As soon as the profits resulting from an employment of money, whatever it may be, increase or diminish, capitals turn in that direction or withdraw from other employments, or withdraw and turn towards other employments, and this necessarily alters in each of these employments, the relation between the capital and the annual product. (1766a, p. 87)

This investment analysis must be seen as a substantial advance on the earlier literature, and hence as a major contribution to economics.

Turgot's other economic writings can be seem as supplementing the analytical framework of the *Reflections*. This can be particularly illustrated from his theory of value, the outlines of which had been developed by the early 1750s (Schelle 1913–23, vol. 1, p. 385). Its foundation rested on a relationship between current (market) price and fundamental value dependent on competition and resources mobility. Subsequently (Turgot, 1767b, p. 120 n.) this proposition is elaborated to demonstrate that the market price of a commodity 'ruled as it is by supply and demand' and liable to 'very sudden fluctuations' though 'not in any essential proportion to the fundamental value, . . . has a tendency to approach it continually, and can never move far away from it permanently'. Turgot therefore developed the classical position which saw 'natural prices' as the centres of gravitation for market prices. Elaborations on the elementary theory of wages of the *Reflections* (1766a, pp. 45–6) are made within this value framework. Turgot did this in a letter to Hume where on the argument that taxes increase 'the fundamental price of labour' or 'the cost of his subsistence', a tax on wages must be rapidly absorbed in market wage rates (Schelle 1913–23, vol. 2, pp. 662–3). The *Reflections* was confined to brief explanations of the current or market price; the unfinished 'Value and Money', with its unsuccessful attempts at determining value in various exchange situations, appears as an elaboration of the underlying competitive theory rather than as a new departure towards a more subjective value theory. Turgot's famous analysis of the 'law of variable proportions' (Schumpeter 1954, pp. 260–1) may also be noted here. This arose in criticism of a common Physiocratic assumption that product was invariably proportional to advances. Turgot (1767b, pp. 111–12) argued instead that as 'advances are gradually increased up to the point where they yield nothing, each increase would be less and less productive', thereby clearly recognizing the possibility of diminishing returns.

On the basis of these contributions, Schumpeter (1954, pp. 260–1, 307, 332) argued the Turgot was a writer in advance of his time by anticipating much of what became important after the 'marginal revolution'. Turgot's analysis of the market mechanism is reminiscent of Böhm-Bawerk and Menger; his 'interest and capital theory . . . clearly foreshadowed much of the best thought in the last decades of the nineteenth century'. However, it seems more reasonable to conclude 'that the resemblance between Turgot's economics and that of post-1870 writers is superficial' and that both in temperament and thrust his economics is part of the classical tradition (Groenewegen 1982). His development of the Physiocratic notion of reproduction (Turgot, 1763; 1766a, pp. 75–6) and his emphasis on the principle of competition as regulator of the rate of interest, wages and values in general, are firmly within that 'classical tradition rehabilitated by Sraffa' (Ravix and Romani 1984, p. 145).

Turgot's strong laissez-faire position, which turned him into the patron saint of the French liberal economics tradition of the middle of the 19th century, was most systematically expounded in his eulogy of Gournay (1759). This rested on the principle that unrestrained self-interest yields the best results in economic activity, a principle he applied wherever he could during his administrative career. It justified his pleas (1770b) and subsequent imposition of domestic free trade in grain, his criticism of the prevalent regulation of lending at interest (1770a), and the suppression of the guilds in one of his famous 1776 Edicts. More important is his discussion of taxation principles. Turgot's major paper on the subject (Turgot, 1763), after setting out some general principles, defends the concept of the single tax on net product on the basis of Physiocratic theory. However, he identifies difficulties in its implementation. These need detailed examination if the benefits of the policy are to be achieved. More generally, it can be said of his policy implementation that though based on broad principles, these were in practice always modified to cater for actual circumstances.

Turgot's work and its importance in the history of economics have occasionally been vigorously debated, most notably in the controversies over the degree of influence he exerted on Adam Smith

and Böhm-Bawerk's interpretation of his capital and interest theory. An assessment of the evidence (Groenewegen 1969) suggests that Turgot influenced Smith on only a few fairly specific points and that the broad similarities (and differences) in their economic systems are largely explained by their common heritage of British and French predecessors. The quarrel over Böhm-Bawerk's interpretation of Turgot's interest theory (involving Cassel, Wicksell and Marshall) is more instructive for the light it sheds on the participants than for discovering Turgot's views on the subject. For example, it can be suggested that Böhm-Bawerk's position may have been influenced by his considerable youthful debts to Turgot's theory while Marshall's involvement may be explained by antipathy to the Austrian economists and some striking similarities between his and Turgot's interest theory (see Groenewegen 1983b). This debate highlights his analytical contributions to interest and capital theory.

The doctrine of social progress, which played such an important part in establishing Turgot's vision, was also applied by him to his history of ideas. In his fragment *On Universal History* (1751b, pp. 95–6) a cumulative notion of intellectual progress is presented, in which ideas are seen to develop necessarily from the systems of predecessors; each scientist, as it were, standing on the shoulders of those who came before. Two decades later, Turgot applied this doctrine to the history of economics when defending Melon, the financial economist, against Du Pont's charge that Melon's work was historically unimportant because it was wrong. 'Someone entering the world after Montesquieu, Hume, Cantillon, Quesnay, M. de Gournay, etc. is less struck by the merit arising from Melon's priority because he does not appreciate it; for him it is no more than a date, and when he reads him, he knew already more than his book' (Turgot to Caillard, 1 January 1771, in Schelle 1913–23, vol. 3, p. 500). This line of thought can be applied to Turgot himself. He built on the work of Montesquieu, Hume, Cantillon, Quesnay and Gournay, thereby becoming a major participant in constructing 18th-century classical political economy with noteworthy contributions of his own particularly to the theory of value, capital and interest, production and distribution.

Turgot's works were collected on three occasions: by his friend Du Pont (Turgot, 1808–11), by Daire and Dussard (Turgot, 1844), and by Schelle (1913–23) together with a biography and associated material. Few of his writings were published in his lifetime, but from 1788 to 1792 some of his major economic writings were republished by his friends Condorcet and Du Pont. Comparison of these texts, manuscript versions and the text of the collected works suggests differences attributable to Du Pont, who edited the text for ideological and occasionally political reasons (see Groenewegen 1977, pp. xxxiv–xxxvi). Schelle first drew attention to, and then removed, many of these corrections, but was not completely successful in this. For this reason, and because of its omissions, particularly of subsequently discovered items from Turgot's voluminous correspondence, Schelle's edition can no longer be described as definitive. Preparing such an edition of Turgot's works awaits both the generous financing required for the task and the services of a devoted editor.

See Also

- ▶ Physiocracy
- ▶ Smith, Adam (1723–1790)

Selected Works

1749. *Letter to M. l'abbé de Cicé on the replacing of Money by Paper, also known as the 'Letter on paper money'*. Trans. Groenewegen (1977).

1750a. *Philosophical review of the successive advances of the human mind*. Trans. Meek (1973).

1750b. Remarques critiques sur les Réflexions Philosophiques de Maupertuis *Sur l'Origine des langues et la signification des mots.* In Schelle (1913–23, vol. 1).

1751a. Lettre á Madame de Graffigny sur les *Lettres d'une Péruvienne*. In Schelle (1913–23, vol. 1).

1751b. *On Universal History.* Trans. Meek (1973).
1753–4. *Remarks on the notes to the translation of Josiah Child* [by Gournay]. Trans. Groenewegen (1977).
1755. *Reflections on the expediency of a law for the naturalisation of foreign protestants*, by Josiah Tucker. Trans into French with notes by Turgot. London\Paris.
1756. *Etymologie, Existence, Expansibilité.* In *Encyclopédie ou Dictionnaire Raisonné des Sciences, des Arts, et Des Métiers*, vol. 6. Paris.
1757a. *Fairs and markets*. Trans. Groenewegen (1977).
1757b. *Foundations*. Trans. as Article I in the Appendix to Condorcet, *Life of M. Turgot*, London, 1787.
1759. *In praise of Gournay.* Trans. Groenewegen (1977).
1760. *Salomon Gessner, la Mort d'Abel, Poéme*. Traduit par M. Huber [et Turgot]. Paris.
1762. *Salomon Gessner, Idylles et Poémes Champêtres.* Traduit par M. Huber [et Turgot]. Lyon.
1763. *Plan for a paper on taxation in general.* Trans. Groenewegen (1977).
1766a. *Reflections on the production and distribution of wealth*. Trans. Groenewegen (1977).
1766b. *On the Characteristics of La Grande and La Petite Culture.* In *Quesnay, Farmers 1756 and Turgot, Sur la Grande et la petitite Culture*, ed. P. Groenewegen Reprints of Economic Classics, Series 2, No. 2, Sydney: University of Sydney, 1983.
1767a. Extract d'un mémoire de M.C. qui contient les principes de l'administration politique, sur la propriété des carrières et des mines, et sur les règles de leur exploitation. In *Ephémérides du Citoyen*, vol. 7. Paris.
1767b. *Observations on a paper by Saint-Péravy.* Trans. Groenewegen (1977).
1767c. *Observations on a paper by Graslin*. Trans. Groenewegen (1977).
1769. Value and money. Trans. Groenewegen (1977).
1770a. *Letters on the grain trade*. Extracts Trans. Groenewegen (1977).
1770b. *Paper on lending at interest*. Extracts trans. in Groenewegen (1977).
1773. *Letter to l'abbé Ternay on the 'Marque des Fers'.* Trans. Groenewegen (1977).
1778. *Virgile, Didon, Poéme en vers métrieque hexamètres, divisés en trois chants, traduit du 4e livre de l'Enéide*. n.p.
1808–11. *Oeuvres de Turgot précédées et accompagnées de mémoires et de notes sur son vie, son administration et ses ouvrages*, ed. P.S. du Pont de Nemours. Paris.
1844. *Oeuvres de Turgot, nouvelle édition*, ed. E. Daire and H. Dussard. Paris.

Bibliography

Bordes, C. 1981. Présentation. In *Turgot, économiste et administrateur*, ed. C. Bordes and S. Morange. Paris: Presses Universitaires de France.
Bourrinet, J. 1965. Turgot, théoricien de l'individualisme libéral. *Revue d'histoire économique et sociale* 43: 465–489.
Cartelier, J. 1981. La contradiction terre/capital-argent chez Turgot. In *Turgot, économiste et administrateur*, ed. C. Bordes and J. Morange. Paris: Presses Universitaires de France.
Condorcet, J.A.N. Caritat de. 1786. *The life of M. Turgot, controller-general of the finances of France in the years 1774, 1775 and 1776*. Trans. from the French with an Appendix, London, 1787.
Dakin, D. 1939. *Turgot and the Ancien Régime in France*. London: Methuen.
de Hausset, Mme. n.d. *Secret Memoirs of the Court of Louis XV and XVI, taken from the Memoir of Madame of Hausset, Lady's maid to Madame de Pompadour and from the Journal of the Princess Lamballe*. London: Grolier Society.
Du Pont, P.S. 1782. *Mémoires sur la vie et les ouvrages de M. Turgot*. Philadelphia\Paris.
Faure, E. 1961. *La disgrâce de Turgot, 12 Mai 1776*. Paris: Gallimard.
Finzi, R. 1981. Turgot: l'histoire et l'économie: 'Nécessité' de l'économie politique? 'Historicite' des lois économiques? In *Turgot, économiste et administrateur*, ed. C. Bordes and J. Morange. Paris: Presses Universitaires de France.
Finzi, R. 1982. The theory of historical stages in Turgot and Quesnay. *Kenzei, Kenkyu* 33 (2): 109–118.
Groenewegen, P.D. 1969. Turgot and Adam Smith. *Scottish Journal of Political Economy* 16 (3): 271–287.
Groenewegen, P.D. 1971. A re-interpretation of Turgot's theory of capital and interest. *Economic Journal* 81: 327–340.
Groenewegen, P.D. 1977. *The economics of A.R.J. Turgot*. The Hague: Martinus Nijhoff.

Groenewegen, P.D. 1982. Turgot: Forerunner of neo-classical economics? *Kenzei Kenkyu* 33 (2): 119–133.
Groenewegen, P.D. 1983a. Turgot, Beccaria and Smith. In *Italian Economics, Past and Present*, ed. P. Groenewegen and J. Halevi. Sydney: Frederick May Foundation.
Groenewegen, P.D. 1983b. Turgot's place in the history of economic thought: A bicentenary estimate. *History of Political Economy* 15: 585–616.
Meek, R.L., ed. 1973. *Turgot on progress, sociology and economics*. Cambridge: Cambridge University Press.
Meek, R.L. 1977. Smith, Turgot and the four stages theory. In *Smith, Marx and after*, ed. R.L. Meek. London: Chapman & Hall.
Morley, J. 1886. Turgot. In *Critical Miscellanies*, vol. 2. London: Macmillan.
Nogaro, B. 1944. *Le développement de la pensée économique*. Paris: Pichon and Durand-Auzias.
Ravix, J.T., and P.M. Romani. 1984. Argent, 'Capital' et reproduction chez Turgot. In *Production, circulation et monnaie*, ed. R. Arena et al. Paris: Presses Universitaires de France.
Schelle, G. 1913. *Oeuvres de Turgot et documents le concernant*. Paris: Félix Alcan.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press\London: Allen & Unwin.
Tsuda, T. 1974. *Catalogue des livres de la bibliothèque de Turgot*. Tokyo: Hitotsubashi University.

Turnpike Theory

Lionel W. McKenzie

Abstract

This account of turnpike theorems concentrates on the discrete time model, descended from the early von Neumann growth model and the Dosso model. It portrays the current state of the theory under the following five headings: (i) a turnpike in the von Neumann model, (ii) a turnpike in the Ramsey model, (iii) Ramsey models with discounting, (iv) turnpike theorems for competitive equilibria, and (v) further generalizations. It emphasizes von Neumann facets and neighborhood convergence as the author's principal contribution to the theory. Under (v), it discusses models that allow for habit formation so that current preferences are affected by past consumption, and for non-convex technologies that have an initial phase of increasing returns followed by a terminal phase of decreasing returns. The theorems that have been reviewed are all concerned with the convergence of optimal paths to stationary optimal paths. However, the method of the proofs is to show that optimal paths converge to one another. The considerable literature on continuous time models related to the literature on the investment of the firm and to the engineering literature on optimal control, as well as applications of the asymptotic results of optimal growth theory to the theory of finance, have not been reviewed.

Keywords

von Neumann growth model; Ramsey model; Asymptotic convergence; Neighborhood turnpike theorem; Competitive equilibrium; Intertemporal resource allocation

JEL Classifications

C62; D90; Q23

The classical economists discussed the eventual convergence of the economy to a stationary state as a consequence of the growth of population and the accumulation of capital, in the absence of continual technical progress or continual expansion of natural resources (Mill 1848, Book IV, ch. V). In their theories they proposed a natural level of real wages equal to subsistence wages and a natural rate of profit just sufficient to prevent decumulation of capital. With a given amount of land and given methods of production, wages and profits would tend to these levels. Both the classical and the neoclassical economists also described a progressive state of the economy where capital accumulates faster than population grows and where technical improvements occur. Cassel (1918, ch. 1, section 6) explicitly considers a 'uniformly progressing state' in which resources and population grow at the same constant rate. However, there is no suggestion that competitive equilibrium converges to such a state.

Ramsey (1928) introduced another type of convergence result in which population, natural resources, and technology are constant but capital is accumulated in an optimal way, that is, in a way to maximize in some sense the sum of utility from consumption over the future. In an aggregative model with one good he describes an optimal path for the capital stock that converges over time to the stock providing the maximum sustainable utility.

A second development in capital theory occurred a few years later and provided the second component of the eventual asymptotic theory for optimal paths. A disaggregated model of capital accumulation was described by von Neumann (1937). In this model there were many alternative production processes with many capital goods as inputs and as outputs. However, labour inputs and consumption did not appear explicitly. In effect, labour was treated as an intermediate product produced by given consumption processes, which were integrated into the processes presented in the model. These processes had stocks of goods as their only inputs and outputs, so all flows of services and intermediate products were suppressed through integration with other processes. Also there were no scarce non-producible goods (natural resources). In this model with a finite number of goods and processes von Neumann proved that there exists a kind of competitive equilibrium in which the maximal rate of uniform expansion of capital stocks is achieved. He proved that this equilibrium is supported by prices in the sense that activities in use earn zero profits and other activities earn zero or negative profits. Also the interest rate implicit in the price system is equal to the maximum rate of expansion.

A Turnpike in the von Neumann Model

The von Neumann model may be defined by an input matrix $A = [a_{ij}]$ and an output matrix $B = [b_{ij}]$. The term $a_{ij} \geq 0$ represents the input of the ith good needed at a unit level of the jth activity, and $b_{ij} \geq 0$ represents the output of the ith good achieved at a unit level of the jth activity. There are n goods and m activities so A and B are $n \times$ m. Inputs occur at the start of a production period, which is uniform for all activities, and outputs appear at the end of the period. Goods at different levels of depreciation are treated as different goods but the number of goods may be as large as needed to achieve an adequate approximation to reality.

An equilibrium of the von Neumann model is defined by a price vector $p \geq 0$, a vector of activity levels $x \geq 0$, and a rate of expansion $a > 0$ which satisfy the relations (1) $Bx \geq \alpha Ax$, (2) $pB \leq \alpha pA$, and (3) $pBx > 0$. Relation (1) provides that output is adequate to supply next period's input requirements. Relation (2) implies that no activity is profitable. Relation (3) implies that some good that is produced has a positive price. Since the relations are homogeneous in x and p, x and p may be chosen to satisfy $\sum_1^m x_j = 1$ and $\sum_1^n p_t = 1$. If (1) is multiplied on the left by p and (2) on the right by x, we find that $pBx = \alpha pAx > 0$. Therefore, some activities are used and they earn zero profits.

Assume the conditions (1) $a_{ij} > 0$ for some i and any j, (2) $b_{ij} > 0$ for some j for any i and (3) if α' is the maximum value of a such that $Bx \geq Ax$ holds for some $x \geq 0$, then $Bx > 0$ (irreducibility). With these conditions (essentially) von Neumann proved that the model has a unique equilibrium, after normalizing x and p, and that the equilibrium value of α is the maximal rate of proportional expansion.

The turnpike name was applied to an asymptotic result for the von Neumann model by Dorfman et al. (Dosso) (1958). They consider paths of accumulation starting from given initial stocks which maximize the size of terminal stocks at the end of the period of accumulation where the proportions of goods in the terminal stocks is specified in advance. They show that for sufficiently long paths which are maximal in this sense the configuration of stocks will be within an arbitrary neighbourhood of the von Neumann equilibrium for all but an arbitrary fraction of the time. This theorem gives the von Neumann equilibrium, which is called 'the turnpike', a general significance for efficient accumulation. An efficient path may be supported by prices just as the equilibrium path is. The turnpike theorem

was conjectured by Samuelson (1966) in an unpublished Rand research memorandum as early as 1949. The Dosso theorem is a local result which was proved (not quite completely) for a two sector model. It was extended in a rigorous way to an n sector model by McKenzie (1963).

A global turnpike theorem was proved for a von Neumann model with many capital goods by Radner (1961), who also introduced the 'value loss' method of proof. This method of proof has been very productive of other turnpike results in subsequent years. Radner's model also allows an infinity of processes and joint production. The equilibrium theorem was extended to this context by Gale (1956). We will consider Radner's theorem in the model with a finite list of processes. Make a further assumption, (4) if (x, p, α) is a von Neumann equilibrium and x^1 0 is any other vector of activity levels, $p(B - \alpha A)x^1 < 0$. With this assumption Radner proved a 'value loss' lemma which may be stated in this way, for any $\varepsilon > 0$ there is $\delta < 1$ such that $pBx^1 \leq (\alpha pAx^1)$, if x_1 is any vector of activity levels, and $|x^1/|x^1| - x/|x|| > \varepsilon$. With this lemma it is easy to prove a turnpike theorem.

Let a sequence of captial stock vectors, $(y_0, y_1, \ldots, y_T)$ be a path if there is a corresponding sequence of activity vectors $(x_1, x_2, \ldots, x_T)$ such that $y_t = Ax_{t+1}$ for $t = 0, \ldots, T-1$, and $y_t \geq Bx_t$ for $t = 1, \ldots, T$. Assume that the vector y_0 of initial stocks satisfies $y_0 > 0$. Then by disposal $y < y_0$ may be chosen so that $y = Ax$ and (x, p, α) is a von Neumann equilibrium. Then $(y, \alpha y, \ldots, \alpha^T y)$ is a feasible path (the comparison path). Suppose $(y_0, y_1, \ldots, y_T)$ is a maximal path. Then the value loss lemma implies that for any $\varepsilon > 0$ there is $\delta < 1$ such that $\delta\alpha py_t \geq py_{t+1}$when $|x_t/|x_t| - x/|x|| > \varepsilon$. But the equilibrium conditions imply $\alpha py_t \geq py_{t+1}$. Thus if $x_t/|x_t|$ is outside the ε-neighbourhood of $x/|x|$ for τ periods then for $T > \tau$ it will be true that

$$\delta^\tau \alpha^T py_0 \geq py_T. \tag{1}$$

Let the desired configuration of terminal stocks be given by the vector y. Define a utility function on terminal stocks by $\rho(z) = \min\ z(i)/\overline{y}(i)$ over $i = 1, \ldots, n$. Then $(y_0, \ldots, y_T)$ maximal implies

$$\rho(y_T) \geq \rho\left(\alpha^T y\right) = \alpha^T \rho(y). \tag{2}$$

Since $y > 0$, $\rho(y) > 0$. Now choose the length of the equilibrium price vector p so that $p(i) \geq 1/\overline{y}$ (i) for some i with $p(i) > 0$. This implies that

$$pz \geq z(i)/\overline{y}(i) \geq \rho(z). \tag{3}$$

Combining (1), (2), and (3), gives the sequence of inequalities

$$\alpha^T \rho(y) \leq \rho(y_T) \leq py_T \leq \delta^T \alpha^T py_0 \tag{4}$$

The first and last terms of (4) imply that δ^τ cannot exceed $\rho(y)/py_0$ which is a well defined positive number. Thus (4) implies that an integer $\overline{\tau}$ exists such that $x_t/|x_t|$ cannot lie outside the ε-neighbourhood of $x/|x|$ for more than $\overline{\tau}$ periods regardless of the length T of the accumulation path. Since y and y_t are linear transforms of x and x_{t+1}, an analogous statement holds for y and y_t. This is a stronger form of the conclusion of the original Dosso theorem.

A Turnpike in the Ramsey Model

Ramsey's aggregative model of capital accumulation was extended to a model with a growing population by Koopmans (1965). Koopmans defines optimality as the maximization of a sum of per capita utilities. When utility is not discounted, he proves that the optimal path converges monotonically to the stock that provides maximum sustainable per capita utility. However, he is also able to treat the case of discounted utility, a case which was not analysed in a satisfactory way by Ramsey. The optimal path of stocks in the discounted case is shown to converge monotonically to the stock for which the marginal product of capital is equal to the sum of the rate of population growth and the rate of discount on utility. The generalization of this result to the many goods case proved very difficult and was only achieved much later by Scheinkman (1976) and Cass and Shell (1976). Also Cass (1966) proved a turnpike theorem for this discounted model with a per capita objective in the sense of Dosso where the accumulation period is finite and a terminal stock is specified. This was proved in an aggregative model.

The spirit of the original turnpike theorem is not well preserved in the aggregative model since the emphasis in the original theorem lies on the relative composition of the capital stock. Samuelson and Solow (1956) generalized the original Ramsey analysis to many goods in a model based on a strictly concave social production function. However, the first rigorous proof of a turnpike result in a Ramsey setting with growing population and more than one capital good was given by Atsumi (1965) in a neoclassical model with two goods, the Dosso model with a Ramsey style objective stated in terms of utility sums. On analogy with the theorem of von Neumann, Atsumi established the existence of a unique maximal balanced growth path along which capital stocks expand at the rate of population growth. The path is maximal in the sense that per capita consumption is maximized over the set of balanced growth paths expanding at the rate of population growth. However, the path is also the only such balanced growth path that is efficient. It is price supported like the von Neumann path, except that consumption goods are now treated as net output rather than as intermediate product. The rate of interest is equal to the growth rate as in von Neumann's case. If the growth rate of population is zero the balanced growth path represents capital saturation, or Ramsey's bliss.

In order to prove the turnpike theorem Atsumi proved a value loss lemma analogous to Radner's lemma for the von Neumann model. He also gave a new definition of an optimal path, that a path is optimal if its utility sums over sufficiently longer initial periods exceed the utility sums of any given alternative path over the same periods. This criterion, in variant forms, is called the overtaking criterion. It was proposed in the same journal issue in a slightly different form, allowing 'eventually equalling' as well as 'exceeding', by Weiszäcker (1965), who used it to discuss the existence of optimal paths. Atsumi proves that infinite optimal paths converge to the maximal balanced growth path. He also proves a theorem for finite optimal paths with assigned terminal stocks analogous to the Dosso theorem.

Turnpike theorems for the general multisector model with a Ramsey objective and a von Neumann technology were first proved by Gale (1967) and McKenzie (1968). Their order of proof does not differ from that of Atsumi, which is, in turn, parallel to the proof used by Radner in the model with maximal growth as an objective. It is simplest to stay close to to the von Neumann model. Let Y represent a reduced model for the activity of one period where $(u, y, -x)$ is a typical element of Y. In the typical element, u is a real number giving a per capita utility level for the period, $y \leq 0$ is a vector of terminal stocks, and $x \geq 0$ is a vector of initial stocks. There are n goods. Since Y is independent of time and of past activities, this model does not allow depletable natural resources or technical progress. Assume

I. Y is a closed convex subset of $2n + 1$ dimensional Euclidean space. Also $(u, y, -x) \in Y$ implies that $(u', y', -x') \in Y$ when $u' \leq u$, $0 \leq y' < y$, and $x' \leq x$, that is, there is free disposal.
II. There is $\overline{x}$ such that $(u, y - \overline{x}) \in Y$ and $y > \overline{x}$, that is, $\overline{x}$ is expansible.
III. (a) For any ξ there is η such that $|x| < \xi$ and $(u, y, -x) \in Y$ implies $u < \eta$ and $|y| < \eta$. (The output from bounded input is bounded.) (b) There is ζ and $\gamma < 1$ such that $(u, y, -x) \in Y$ and $|x| \geq \zeta$ implies $|y| < \gamma|x|$. (All paths are bounded.)
IV. If $(u^*, k^*, -k^*) \in Y$ and $u^* \geq u$ for any $(u, y, -x) \in Y$, then k^* is expansible.

A path in this model is a sequence $(u_t, k_t, -k_{t-1}) \in Y$ for $t = 1, \ldots, T$, with T finite or infinite and $(u_t, k_t, -k_{t-1}) \in Y$ for each t. The existence of k^* which defines a path of maximal utility at balanced growth is guaranteed by assumptions I, II, and III. It may be proved that a price vector $p^* \geq 0$ exists which satisfies the support property.

$$u + p^*(y - x) \leq u^* + p^*(k^* - k^*) = u^*, \quad \text{for any } (u, y, -x) \in Y. \tag{5}$$

Define the von Neumann facet F of the technology set Y as all $(u, y, -x) \in Y$ such that $u + p^*(y - x) = u^*$. This means there will be no value loss when the path of accumulation lies on F.

However, when the path is off F, a value loss lemma due to Atsumi (1965) applies and for any $\varepsilon > 0$ there is $\delta > 0$ such that if $(u, y, -x)$ lies outside the $\in$-neighbourhood of F it follows that

$$u + p^*(y - x) \leq u^* - \delta. \tag{6}$$

As before a comparison path is found, but no longer simply by disposal to an equilibrium. a lemma due to Gale (1967) implies that a path exists from any expansible stock to any other expansible stock. Thus k_0 expansible and k^* expansible (Assumption IV) implies that a path exists from k_0 to $k_s = k^*$ for some integer s. Then $k_t = k^*$ may be maintained indefinitely for $t > s$.

With this preparation the proof of the turnpike theorem as given by McKenzie (1968) is straightforward. Let u_1 be the utility sum along the approach to k^* from k_0 along the comparison path and let $(u_t, k_t, -k_{t-1}\ t = 1, 2, \ldots$ be an optimal path by the overtaking criterion. Suppose the optimal path lies outside the ε-neighbourhood of F for τ periods. Let u_1 be the period in which the optimal path overtakes the comparison path and let $T > \bar{t}$ be chosen arbitrarily. Consider the inequalities

$$\begin{aligned}(T - s)u^* + u_1 &\leq \sum_1^T u_t \leq Tu^* \\ &\quad - \sum_1^T p^*(k_t - k_{t-1}) - \tau\delta \\ &= Tu^* - p^*(k_T - k_0) - \tau\delta. \end{aligned} \tag{7}$$

The first inequality is justified by $T > \bar{t}$ and optimality. The second inequality follows from (5) and (6). However, (7) implies that $\tau \leq (su^* - u_1)/p^*(k_t - k_0)$, which is a constant. Since $\in$ is arbitrary, it is clear that an optimal path converges to F. Indeed, as one might suspect from (7), paths that are any good converge to F, since paths that do not converge become indefinitely worse than the comparison path. Also there is no difficulty in proving the analogue of the Dosso type of theorem for finite paths when terminal stocks are specified.

In order to prove that optimal paths converge to a maximal balanced path or, in per capita terms, a maximal stationary path, the assumptions must be strengthened. The most general assumption is that the von Neumann facet F is stable, in the sense that all paths that remain on the facet forever converge uniformly to a maximal stationary path. The assumption in this general form was proposed by Inada (1964) for the von Neumann model. Then it may be proved that a path which converges to a stable facet must also converge to the stable point on the facet (see McKenzie 1968). The simplest case arises when $F = \{(u^*, k^*, -k^*)\}$. This is analogous to the case analysed by Radner for the von Neumann model. It is implied if it is assumed that the technology set Y is strictly convex. However, it should be noted that strict convexity of Y is not consistent with neoclassical models of production when goods are produced by independent industries, even though consumer utilities are strictly concave functions of consumption. Indeed, as Bewley (1982) has pointed out, it is not consistent with the use of machines in production since an input of m machines leads to an output of m older machines at the end of the period. The behaviour of paths on F may be studied by means of difference equations which are defined in terms of points $(u_j, y_j, -x_j) \in F$ that span F. In the analogous problem for the von Neumann case this was done explicitly for the generalized Leontief model by McKenzie (1963). However, the time that can be spent outside an $\in$-neighbourhood of the turnpike by an optimal path is no longer a given number of periods but a given fraction of the total time of accumulation. Even though the facet is not stable, it was proved by Brock (1970) that if the maximal stationary path is unique the average capital stock of an optimal path over time converges to the capital stock of the maximal stationary path.

Ramsey Models with Discounting

Turnpike theorems for Ramsey models with von Neumann technologies and positive discounting of utility are much harder to prove. The difficulty is that discounting utility implies discounting value losses, so value loss is bounded and need not exceed the loss from going over to a comparison path. The first theorems were due to

Scheinkman (1976) and Cass and Shell (1976). These were global results proved for models defined by differentiable functions. Also the theorems are proved for discount factors sufficiently close to 1. Scheinkman showed for discount factors close to 1 and with Y strictly convex that the optimal path would visit a small neighbourhood of the maximal stationary path at least once. Then he showed that if this neighbourhood were small enough the path would not leave it but would in fact converge to the maximal stationary path.

The reduced model that is frequently used for the multisector Ramsey case with discounting expresses per capita utility over a period by a function $u(x, y)$ where x is the vector of initial stocks per capita and y is the vector of terminal stocks per capita. The function expresses the maximum utility achievable during the period given these and conditions. It is assumed that utility in one period is independent of events that occur in other periods except as they influence the initial or terminal stocks of that period. The technology set Y described for the model without discounting corresponds to the epigraph of the function u. The function u is defined on a convex set D contained in the positive orthant of a Euclidean space of dimension $2n$, the Cartesian product of the space of initial stocks and the space of terminal stocks. Let the discount factor be $\rho < 1$. The assumptions are

I′. The utility function $u(x, y)$ is concave and upper semi-continuous. If $(x, y) \in D$, then $(z, w) \in D$ for all $z \geq x$ and all w such that $0 \leq w \leq y$. Also $u(z, w) \leq u(x, y)$ holds (free disposal).

II′. There is $\bar{x}$ such that $(\bar{x}y \in D)$ and $\rho y > \bar{x}$, that is $\bar{x}$ is ρ-expansible.

III′. For any ξ there is η such that $|x| \leq \xi$ and $(x, y) \in D$ implies $|y| < \eta$. (b) There is ζ and $\gamma < 1$, such that $(x, y) \in D$ and $|x| \geq \xi$ implies $|y| < \gamma|x|$.

These assumptions are closely parallel to the assumptions of the undiscounted Ramsey model except that the expansibility assumption is strengthened. It is easily seen that any neoclassical model with utility defined on consumption and labour services and a production function that converts inputs of labour services, initial stocks of capital, and natural resource flows into outputs of capital goods and consumption goods corresponds uniquely to a reduced model. The utility function of the reduced model is derived by maximizing the utility of consumption and labour services given the initial and terminal stocks of capital goods. Upper semi-continuity of utility carries over from the neoclassical model to the reduced model.

A path is a sequence $\{k_t\}$, $t = 1,2,\ldots$, such that $(k_{t-1}, k_t) \in D$ for all t.

A path $\{k_t\}$ is optimal if $\sum_1^\infty \rho^t u(k_{t-1}, k_t) \geq \sum_1^\infty \rho^t u(k'_{t-1}, k'_t)$ holds for every path $\{k'_t\}$. On Assumptions I′, II′, and III′ it may be proved that a stationary path with $k_t = k$, all t, exists which is price supported in the sense that there exists a vector $q \geq 0$ such that

$$u(z, w) + qw - \rho^{-1}qz \leq u(k, k) \\ +qk - \rho^{-1}qk \text{ for all } (z, w) \in D. \tag{8}$$

We may use the relation (8) to prove that $k_t = k$ defines an optimal path from k. Let k'_t be any path from k with $k'_0 = k$. Then summing the relations (8) over the path gives for any $T \geq 1$,

$$\sum_1^T \left[\rho^1 u(k'_{t-1}, k'_t) - u(k, k)\right] \\ = \rho^T q(k - k'_T) - \sum_1^T \delta_t. \tag{9}$$

where $\delta_t \geq 0$. Since k'_t is non-negative and bounded, and $\rho^T \to 0$ as $T \to \infty$, in the limit the right-hand side of (9) is less than or equal to 0 which establishes the optimality of $\{k_t\}$.

In the discounted model it is also useful to have prices to support any optimal path. Let $V(x) = \sup \sum_1^\infty \rho^t u(k_{t-1}, k_t)$ over all paths $\{k_t\}$, $t = 0, 1, \ldots$, with $k_0 = x$. Put $V(x) = -\infty$ if no path from x exists. Define a sufficient stock as a stock from which there exists a path that reaches an expansible stock in finite time. It has been proved by McKenzie (1974), extending a result of Weitzman

(1973), that a price sequence $\{q_1\}$ exists for any optimal path $\{k_t\}$ that starts from a sufficient stocks k_0. This price sequence satisfies

$$u(k_t, k_{t+1}) + q_{t+1}k_{t+1} - \rho^{-1}q_t k_t \geq u(x,y) + q_{t+1}y - \rho^{-1}q_t x \quad \text{for all} (x,y) \in D, \tag{10}$$

and

$$V(k_{t+1}) - q_{t+1}k_{t+1} \geq V(y) - q_{t+1}y, \quad \text{for all } y \text{ such that } V(y) > -\infty \text{ for } t > \tau, \tag{11}$$

where τ is independent of ρ.

Using the prices q_1 as well as the prices $q(\rho)$ that support a stationary optimal path $k(\rho)$ allows the definition of a symmetric value loss function $L(t) = [q_t(\rho) - q(\rho)\,(k_t(\rho) - k(\rho)]$ which can play a crucial role in proving turnpike theorems for this model. It should be kept in mind that $L(t)$ depends on ρ and on the particular $k(\rho)$ as well.

If one writes (10) with $(x, y) = [k(\rho), k(\rho)]$ and then writes (10) again with the roles of k_t and $k(\rho)$ reversed, with the support prices of $k(\rho)$, that is, $[q(\rho), q(\rho)]$ in place of (q_t, q_{t+1}), subtracting the second version of (10) from the first gives the result

$$L(t+1) - \rho^{-1}L(t) \geq 0. \tag{12}$$

Doing the same operation with (11) gives

$$L(t) \leq 0. \tag{13}$$

Then using $L(t)$ in a similar manner to the asymmetric value loss for the undiscounted model a turnpike result may be proved for paths which start from a sufficient stock.

Define a von Neumann facet $F[k(\rho)]$ for this model as the set of all $(x, y) \in D$ such that $u(x, y) + q(\rho)y - \rho^{-1}\,q(\rho)x = u[k(\rho), k(\rho)] + (1 - \rho^{-1})\,q(\rho)k(\rho)$. This is the projection of a flat in the graph of the function $u(x, y)$ on the commodity space. There is a von Neumann facet for every non-trivial stationary optimal path, that is, for every stationary optimal path $k(\rho)$ which satisfies the condition that $u[k(\rho), k(\rho)]$ is maximal over the set of (x, y) such that $\rho y - x \geq (\rho - 1)k$.

If $\rho = 1$, this set is the same as that over which $u(k, k)$ is maximal.

Say that a point $(x, y) \in D$ is supported by the prices (p, q) if $u(x, y) + qy - \rho^{-1}\,px \geq u(z, w) + qw - \rho^{-1}\,pz$ for any $(z, w) \in D$.

Two additional assumptions are made. Let $\overline{\rho}$ be a value of ρ for which D satisfies assumption II′. Assume

IV′. If (p, q) are support prices for some point of the von Neumann facet $F[k(\rho)]$ where $\overline{\rho} \leq \rho \leq 1$, then $(p, q) = [q(\rho), q(\rho)]$.

In other words, the von Neumann facet has a unique support. Assumption IV′ implies that $L(t)$ is zero $F[k(\rho)]$ and that $L(t)$ is close to zero in a small neighbourhood of $F[k(\rho)]$. This assumption is needed since $k_t(\rho)]$ is need not be near $k(\rho)$ even though $[k_t(\rho), k_{t+1}]$ is near $F[k(\rho)]$.

The second assumption is needed to obtain a uniform value loss condition which is analogous to (6) in the undiscounted model for all $F(k(\rho))$, $\overline{\rho} \leq \rho < 1$. Let $\delta(k(\rho), (x, y))$ be the deficiency of the right-hand side of (10) when $k_t = k_{t+1} = k(\rho)$ and $q_t = q_{t+1} = q(\rho)$. Assume

V′. For any $\varepsilon > 0$ there is $\delta > 0$ such that $|x| < \xi$ and (x, y) outside the $\in$-neighbourhood of $F[k(\rho)]$ implies that $\delta[k(\rho), (x, y)] > \delta$ for any ρ with $\overline{\rho} \leq \rho < 1$ and any choice of $k(\rho)$.

Rewrite (12) in the form

$$L(t+1) - L(t) \geq \left(\rho^{-1} - 1\right)L(t) + \delta. \tag{14}$$

As a consequence of V′ the value loss δ may be chosen uniformly over ρ and $k(\rho)$. Then ρ may be chosen so that $(\rho^{-1} - 1)L(0) \geq -\delta/2$, uniformly. Moreover, this condition continues to hold for $t \geq 0$ which implies that $L(t)$ grows by an indefinite amount unless the path enters the ε-neighbourhood of $F(k(\rho))$. This it must do or violate (13). Then using Assumption IV′ the existence may be established of an ε'-neighbourhood to which the path is subsequently confined. Also the ε'-neighbourhood can be made arbitrarily small by the choice of ε. For the details of this argument see McKenzie (1983). It is not implied that the $k(\rho)$ are unique or that cyclic paths, or even chaotic paths,

are absent. However, these paths must eventually lie in the ε'-neighbourhood. The larger is the neighbourhood chosen, the smaller the discount factor that may be allowed. It is implied that all $k(\rho)$ for a given ρ lie in the ε'-neighbourhood of any one of the $F(k(\rho))$. An example using the discounted utility function $\rho^t u(x,y) = \rho^t x^\beta (1-y)^{1-\beta}$, $0 < \rho \leq 1$, $0 < \beta < 1$ is described in McKenzie (1983).

As in the undiscounted model it is possible to go beyond facet stability on further assumptions. Most simply if strict concavity of $u(x, y)$ is assumed the von Neumann facets are trivial since $F[k(\rho)] = \{\{k(\rho)\}\}$. Then a neighbourhood theorem is implied for the nontrivial optimal stationary paths, each of which must lie within e of any other. Indeed, it may be shown for differentiable $u(x, y)$ that the optimal stationary path is unique for ρ near enough to 1 (Brock 1973). Indeed, Boldrin and Montrucchio (1986b) have a simple condition on ρ in terms of the concavity of u which implies uniqueness. Even without strict concavity convergence to a neighbourhood of $k(\rho)$ may be proved if it is assumed that $k(1) = k^*$ is unique (in any case the set of maximal stationary paths is convex) and that the von Neumann facet $F(k^*) = F$ is stable in the sense that all paths in F converge uniformly to (k^*, k^*).

Suppose $u(x, y)$ is strictly concave and twice continuously differentiable. Also assume that the Hessian matrix of u is negative definite at (k^*, k^*). Then ρ may be chosen close enough to 1 so that the matrix

$$Q(\rho) = \begin{bmatrix} \rho u_{xx} & \rho u_{xy} \\ u_{yx} & u_{yy} \end{bmatrix}$$

evaluated at $(k(\rho), k(\rho))$ is negative definite. Here $u_{xy} = \partial^2 u(x, y)/\partial x \partial y$, and analogously for the other blocks for $Q(\rho)$. If k_0 is sufficient, the neighbourhood turnpike theorem implies that the path is eventually confined to a neighbourhood where $Q(\rho)$ is negative definite when ρ is chosen sufficiently near 1. If this neighbourhood is small enough it may be shown that $L(t + 1) - L(t)$ is positive and $L(t) \leq 0$ will be violated unless $k(t)$ converges to k^* (McKenzie 1985). The condition $Q(\rho)$ negative definite assumed over the interior of D was proposed by Brock and Scheinkman (1978) and used to prove a global turnpike theorem.

An asymptotic result is also available when the von Neumann facets have positive dimension. The conditions needed are $Q(\rho)$ negative definite at (k^*, k^*) in directions leading immediately off the facet F, the absence of cyclic paths on F, and the condition that the stable manifold at (k^*, k^*) have a tangent plane whose projection on the input space covers a neighbourhood of k^*. The implications of these conditions for the neoclassical model without joint production have been studied by Takahashi (1985).

The turnpike theorems for the multisector Ramsey model thus far described depend on discount factors near 1. There are some theorems, however, which are free of this condition. Suppose $u(x, y)$ is twice continuously differentiable in the interior of D and strictly concave. Under Assumptions I′–V′ a necessary and sufficient condition for the optimality of a path $\{k_t$ which is bounded away from the boundary of D is

$$u_2(k_{t-1}, k_t) + \rho u_1(k_t, k_{t+1}) = 0, \qquad (15)$$

for $t = 1, 2, \ldots$, Araujo and Scheinkman (1977) consider the Jacobian of the infinite sequence of equations given by (15). They define a notion of dominant diagonal blocks for the Jacobian and show that this, together with local asymptotic stability, implies that k_t converges to $k(\rho)$. However, the assumption of local stability has been shown to be unnecessary (McKenzie 1977). The dominant diagonal block condition is independent of ρ.

An interesting special case arises where $u_{12}[k(\rho), k(\rho)]$ is nonsingular and symmetric and the linearization of (15) at $k_{t-1} = k_t = k_{t+1} = k(\rho)$ does not have roots of unit modulus. Then the dominant diagonal block condition for the optimal stationary path $k_t = k(\rho)$ is necessary and sufficient for local stability (Dasgupta and McKenzie 1985). However, the dominant diagonal block condition is sufficient for global stability, so in this case it is necessary and sufficient for global stability.

Another approach of turnpike theory which is independent of ρ has been found by Boldrin and

Montrucchio (1986a). Assume that $u(x, y)$ is concave and strictly concave in y. Define the binary relation P by yPx if and only if $u(x, x) + \rho V(x) < u(x, y) + \rho V(y)$, where V is the value function. Let the projection of D on the space of initial stocks be a compact set X. P is said to be acyclic if there is no sequence $(x_1, \ldots, x_n)$ such that $x_{i+1}\ Px_i$ for $i = 1, \ldots, n$, when x_{n+1} is set equal to x_1. Define the policy function $f(x)$ equal to the maximizer of $u(x, y) + \rho V(y)$ over y. Then $f(x)Px$ holds whenever $f(x) \neq x$. Any optimal path is generated by repeated applications of f. Suppose the optimal stationary path $k_t = k(\rho)$ is unique. Then if P is acyclic, all optimal paths converge to $k(\rho)$. But it may be seen that P is acyclic if and only if $\sum_1^n u(x_t, x_t) \geq \sum_1^n u(x_t, x_{t+1})$, for every path $(x_1, \ldots, x_n)$ and $x_{n+1} = x_1$. This condition holds, for example, if $u(x, y) = \varphi(x) + \psi(y - x)$, which has been referred to as the separable case. This case occurs in some models of investment by the firm in which u is a profit function (see Treadway 1971).

Turnpike Theorems for Competitive Equilibria

In recent years the circle has been completed and turnpike theorems have led to asymptotic results for the theory of competitive equilibrium. These results bear some analogy to the convergence to a stationary state described by the classical economists. However, the modern results are based on capital accumulation in the absence of a population dynamics. There is also an analogy to the Dosso theorems on efficient capital accumulation in the von Neumann model, but now it is utility maximization rather than efficient capital accumulation that defines the path.

The equilibrium model with many households was described briefly by Ramsey but the first serious analysis of the model was given by Becker (1980), who verified Ramsey's conjecture that the long-run equilibrium would place all capital in the hands of the most thrifty households. Subsequently Bewley (1982) applied the turnpike results of optimal growth theory to prove that the competitive equilibrium path with infinitely lived households, who hold identical discounts on future utility from consumption, will converge to a stationary equilibrium. He made use of the duality of competitive equilibrium and a social optimum which was described by Negishi (1960).

Bewley presents an equilibrium model in which production over the infinite horizon is decentralized to a finite number of firms that possess in each period strictly concave production functions which are twice continuously differentiable. Similarly the consumers, who are finite in number, have strictly monotone, strictly concave, and twice continuously differentiable utility functions in each period. Zero production is possible and primary inputs are necessary for production. Each firm maximizes its profit over the future at present prices. Each consumer maximizes his discounted utility sums from consumption over the future subject to a budget constraint derived from the value of his endowment of primary goods (assumed the same in each period) and his share of firms' profits, all calculated at present prices. To guarantee income each consumer is assumed to have a positive endowment of some primary good which is also a consumption good. It is feasible to produce a positive amount of all goods in all periods. Finally it is assumed that in any stationary equilibrium (with transfer payments) every firm has positive initial stocks of all produced goods. Transfer payments must be allowed in the stationary equilibrium since the consumer's budget is balanced only over the infinite horizon, so that asymptotically he may be (in effect) a debtor or a creditor and (in effect) pay or receive interest.

An equilibrium is a feasible allocation $\{x_h(t), y_f(t)\}$, $t = 1, 2, \ldots$, where h indexes consumers and f indexes firms, together with a price vector $p(t)$ such that $\sum_0^\infty p(t) < \infty$, and where each consumer maximizes utility and each firm maximizes profit. In a stationary equilibrium the budget constraint for consumers is written $px(t) \leq px_h(t)$ to take account of the transfer payments which may occur. It is proved that if all consumers have the same discount factor and this discount factor is sufficiently close to 1, then the allocation of the competitive equilibrium converges exponentially to the allocation of a stationary equilibrium with transfers $\left(\overline{x}_h(t), \overline{y}_f(t), \overline{p}(t)\right)$.

The proof that the equilibrium converges to a stationary state is given by identifying the competitive equilibrium with the maximization of a weighted sum of consumers' utilities which is given by $V(K) = \sup \sum_{t=0}^{\infty} \rho^t \sum_h \Lambda_h^{-1} U_h(x_h(t))$. The weights Λ_h^{-1} are the inverses of the marginal utilities of expenditure, and ρ is the discount factor on future utility, K is the vector of initial stocks of capital (produced goods), V is a maximum of the weighted utility sum over all consumption streams consistent with the initial stocks K, the endowments, and the production functions of the firms. Bewley shows that $V(K(t))$ may be used as a Lyapounov function to establish that $K(t)$ converges to $\overline{K}$ which is the vector of stocks of a stationary optimal path for the optimal growth problem and of a stationary competitive equilibrium with transfers that corresponds to it. The proof depends on the consumers' utility functions being concave, not just quasi-concave. Bewley's theorem differs from the other asymptotic theorems we have considered in that the stationary state depends on the initial stocks since they affect the weights Λ_h^{-1}, that is, the relative wealth of consumers.

Other theorems analogous to Bewley's have been proved by Yano (1984a, 1985). Yano describes a model with a production sector given by a cone with a cross-section which is strictly convex in the neighbourhood of the path achieving maximum sustainable utility. This is analogous to the model used by Radner for the Dosso problem. As a consequence the production sector does not have a multiplicity of firms with distinct technologies since this would lead to flats in the production cone. However, his consumers are allowed to have different tastes and endowments. He dispenses with differentiability and proves a neighbourhood turnpike theorem similar to McKenzie's but applying to a competitive equilibrium path. The turnpike depends on consumer endowments and initial stocks as in Bewley's analysis.

Yano also shows (1984b) that the turnpikes in the discounted model can be brought within an ε-neighbourhood of a stationary competitive equilibrium without discounting for arbitrary $\varepsilon > 0$ for an appropriate choice of ρ and a given initial distribution of capital. In this sense the equilibrium turnpike has an invariance to initial stocks typical of the turnpikes of optimal growth theory. The explanation is that as ρ converges to 1, the importance of initial stocks compared with future endowments becomes negligible in determining the wealth of the consumer. The stationary competitive equilibrium without discounting is based upon the capital stocks that achieve capital saturation, as at Ramsey's bliss point. Thus there is no interest income and by homogeneity of production there is no profit. In other words, consumer wealth is independent of ownership shares in profits (which are zero) or in capital stocks (which have zero rental income). However, the distribution of endowments remains important, both of personal skills and of natural resources.

Some progress has been made toward handling the case of consumers whose discounts on future utility differ without leading to an asymptotic state in which all capital is in the hands of a few patient consumers. To avoid this outcome it is assumed that the discount on future utility depends on the level of utility achieved, either by the consumption of the current period or by the entire future consumption stream. Lucas and Stokey (1984) consider a model with one consumption good and many capital goods. They use a device of Koopmans (1960) to write the utility of a consumption stream ${}_1C = (C_1, C_2, \ldots)$ in terms of the first period consumption and the utility of the stream ${}_2C = (C_2, C_3, \ldots)$, that is, $U({}_1C) = W(C_1, U({}_2C))$. In the case of additively separable utility the formula appears as $U({}_1C) = U(C_1) + \rho U({}_2C)$, where $\rho \leq 1$ is the discount factor. This suggests that the subjective discount factor implied by W be defined by W_2, the derivative of W with respect to its second argument. Consider constant paths ${}_2C = (C, C, \ldots)$. Lucas and Stokey assume, unlike Fisher (1930), that along constant paths the discount factor W_2, is a decreasing function of C. They prove in their model that the stationary state of the perfect foresight economy is unique on this assumption. They also prove a turnpike theorem for the competitive equilibrium of a two consumer model without production.

More recently Epstein (1987), using a continuous time version of the Lucas and Stokey model has succeeded in proving a turnpike theorem for a

model with many consumers whose utility functions and implied discount factors may differ. He defines the implicit rate of time preference as equal to minus the proportional rate of change of marginal utility along a locally constant path. The corresponding periodwise discount factor would be approximately $\rho = 1 - r$. Then r is a function of current consumption and the utility of future consumption. However, he assumes r to be independent of current consumption but increasing in the utility of future consumption. He is able to prove with this assumption that all Pareto optimal paths of accumulation converge to a unique stationary path. Thus by the first welfare theorem all competitive equilibria converge to a unique stationary equilibrium. However, these results depend on implicit rates of time preference sufficiently close to 0, or else own rates of return in production sufficiently close to 0. Unlike the stationary equilibria of Bewley and Yano, this equilibrium is independent of initial stocks and their distribution of ownership. All that matters finally for the wealth of consumers are their endowments, repeated each period, and their rates of time preference. Such a result is compatible with classical ideas. The methods used by Epstein in his proof are derived from the work of Brock and Scheinkman (1976) in a differentiable multisector Ramsey model.

It should be pointed out that the assumptions made by these writers on the function W or U imply a weak form of separability between present and future consumption. Loosely speaking the ranking of future consumption bundles is not affected by current consumption, nor a *fortiori* by past consumption. Thus intuitively what is done by their formulation is to make the rate of time preference, or the implicit discount factor, depend on present consumption and the utility of the future consumption stream.

Further Generalizations

Turnpike theorems have also been generalized to allow habit formation so that current preferences are affected by past consumption (for example, Samuelson 1971; Heal and Ryder 1973). It has been shown by Epstein (1986) that the turnpike theorems hold in these circumstances if and only if a form of asymptotic independence holds or that the effect of earlier consumption on current preferences fades out over time.

The theorems on the correspondence of turnpike theorems for optimal paths and for equilibrium paths have been extended in two ways. Coles (1985) proved asymptotic convergence to the von Neumann facet in a model like that of Yano where separable additive utility is assumed and the discount factors of consumers may differ. This extension allows the use of capital equipment in production without introducing interdependence between industries.

On the other hand, the Bewley theorems were extended by Marimon (1984) to a stochastic model. The preferences and technology, as well as the discount factors and endowments, are made to depend on a stationary and transitive stochastic process. The equilibrium is that of a complete Arrow-Debreu market in which consumers own given shares of the firms and all trading occurs at the initial date. Marimon proves that the equilibrium allocation converges almost surely to the allocation of a stationary equilibrium with transfer payments. This result holds when the discount factors are sufficiently close to 1, almost surely. From the viewpoint of optimal growth his results generalize those of Brock and Mirman (1972) for the one sector model, those of Evstigneev (1974) for the undiscounted multisector model, and those of Brock and Majumdar (1978) for the discounted multisector model.

Conditions have been found to guarantee the presence of cycles in models of economic growth and by the same token in models of competitive equilibrium.

These results are relevant to the study, of endogenous cycles in competitive economies, a study which has been pursued in overlapping generations models by Grandmont (1985). In the context of optimal growth models Benhabib and Nishimura (1985) have given sufficient conditions for robust periodic optimal paths in Ramsey models with additively separable utility and neoclassical technology. The utility function is strictly concave and there is one capital good. If

$u(x, y)$ is the reduced form utility function the basic condition for oscillations on interior optimal paths in the discrete time model is that $u_{12}(x, y) < 0$ hold throughout the interior of D, the domain of definition of u. This says that larger initial stocks (an increase in wealth) cause the marginal utility of terminal stocks to be smaller (saving is discouraged). However, added conditions are needed to ensure sustained oscillations which are robust to small perturbations of the model. In particular, it is assumed there is a ρ such that $u_{22} + \rho u_{11} > (1 + \rho)u_{12}$ where the derivatives are evaluated at a nontrivial stationary optimal path.

This line of research has been further advanced by Boldrin and Montrucchio (1985). Consider a multisector Ramsey model defined by a reduced utility function $u(x, y)$. Assume that $u(x, y)$ is defined over a compact set $D \subset XxX \subset R^n_+ \times R^n_+$ where X is the projection of D on the first factor. Let $u(x, y)$ be continuous and concave, strictly concave in y, strictly increasing in x and strictly decreasing in y. Define the optimal policy function $f(x) = y$ where y is the unique vector of terminal stocks for the first period of an optimal path, when x is the vector of initial stocks. Let θ map X into X and assume that it may be extended to be twice continuously differentiable on x. Then there exists ρ*0 such that, given any ρ with $0 < \rho \leq \rho^*$, θ is the policy function for some Ramsey problem satisfying the usual assumptions. Since the policy function can be chosen freely, it follows that no complex behaviour of optimal paths can be excluded, in particular chaotic paths are possible. Boldrin and Montrucchio also provide a way of calculating a possible value for ρ* in terms of the diameter of X and bounds on the derivatives of θ.

Another direction of generalization is to models with non-convex technologies. The principal turnpike results that have been proved with non-convexity are concerned with one good Ramsey models of optimal growth in which the production function has an initial phase of increasing returns followed by a terminal phase of decreasing returns (for example, Skiba 1978; Majumdar and Mitra 1982; Dechert and Nishimura 1983). Optimal paths in differentiable models with discrete time and discounted utility converge to steady states, that is, stationary paths, among which the origin is included. Nontrivial steady states are solutions of $f'(x) = \rho^{-1}$ where $f(x)$ is a differentiable production function. There cannot be more than two nontrivial steady states, k* in the concave region and k* in the convex region, and there may be none. Every optimal path converges to a steady state. If the discount factor ρ is near 1, optimal paths converge to k*. If ρ is small enough, they converge to 0. For intermediate values of ρ the turnpike depends on the initial capital stock. There is a critical value k_c such that $k_0 < k_c$ implies that optimal paths converge to the origin and $k_0 < k_c$ implies that optimal paths converge to k*. *If* $k_c = k$*, then $k_t = k$*, $t = 0, 1, \ldots$, is the unique optimal path from $k_0 = k_c$.

The theorems that have been reviewed are all concerned with the convergence of optimal paths to stationary optimal paths. However, the method of the proofs is to show that optimal paths converge to one another. Thus it is not really necessary that the reduced utility function be constant over time. Loosely speaking, constancy may be replaced by variation within bounds. The variation of the reduced utility function may reflect a varying production function and a varying utility function over consumption bundles. Examples of this approach are found in Keeler (1972), McKenzie (1974, 1976, 1977), Mitra (1979) and Brock and Magill (1979).

The account of turnpike theorems has concentrated on the discrete time model, which is descended from the early von Neumann growth model and the Dosso model. There is a considerable literature on the continuous time model which is related to the literature on investment in the firm and to the engineering literature on optimal control. With a continuous time model some results become available from the theory of differential equations which permit further theorems to be proved on the asymptotic behaviour of optimal paths. Many of these results are found in Brock and Scheinkman (1977). Particularly complete results for the local problem were found by Magill (1977). Also the asymptotic results of optimal growth theory have been applied in areas

which have not been reviewed, for example, in the theory of finance (see Brock 1982).

This article was first published in *The New Palgrave Dictionary of Economics*, First Edition, 1987 and was updated with an abstract, keywords and JEL codes by Professor M. Ali Khan in June 2012.

See Also

- ▶ Multisector Growth Models
- ▶ Turnpike Theory, A Current Perspective
- ▶ von Neumann Ray

Bibliography

Atsumi, H. 1965. Neoclassical growth and the efficient program of capital accumulation. *Review of Economic Studies* 32: 127–136.

Becker, R.A. 1980. On the long-run steady state in a simple dynamic model of equilibrium with heterogeneous households. *Quarterly Journal of Economics* 94: 375–382.

Benhabib, J., and K. Nishimura. 1985. Competitive equilibrium cycles. *Journal of Economic Theory* 35: 284–306.

Bewley, T.F. 1982. An integration of equilibrium theory and turnpike theory. *Journal of Mathematical Economics* 10: 233–268.

Boldrin, M., and L. Montrucchio. 1986a. On the indeterminacy of capital accumulation paths. *Journal of Economic Theory* 40: 26–39.

Boldrin, M., and L. Montrucchio 1986b. *Acyclicity and stability for intertemporal optimization models*, Working Paper. Rochester: University of Rochester, March.

Boldrin, M. and Montrucchio, L. 1986c. Private communication.

Brock, W.A. 1970. On existence of weakly maximal programmes in a multi-sector economy. *Review of Economic Studies* 37: 275–280.

Brock, W.A. 1973. Some results on the uniqueness of steady states in multisector models of optimum growth when future utilities are discounted. *International Economic Review* 14: 535–559.

Brock, W.A. 1982. Asset prices in a production economy. In *The economics of information and uncertainty*, ed. J.J. McCall. Chicago: University of Chicago Press.

Brock, W.A., and M. Magill. 1979. Dynamics under uncertainty. *Econometrica* 47: 843–868.

Brock, W.A., and M. Majumdar. 1978. Global asymptotic stability results for multisector models of optimal growth with uncertainty when future utilities are discounted. *Journal of Economic Theory* 18: 225–243.

Brock, W.A., and L. Mirman. 1972. Optimal economic growth and uncertainty the discounted case. *Journal of Economic Theory* 4: 479–513.

Brock, W.A., and J. Scheinkman. 1976. The global asymptotic stability of optimal control systems with applications to the theory of economic growth. *Journal of Economic Theory* 12: 164–190.

Brock, W.A., and J. Scheinkman. 1977. The global asymptotic stability of optimal control with applications to dynamic economic theory. In *Applications of control theory to economic analysis*, ed. J.D. Pitchford and Jose A. Scheinkman. Amsterdam: North-Holland.

Brock, W.A., and J. Scheinkman. 1978. On the long-run behavior of a competitive firm. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwödiauer. Dordrecht: D. Reidel.

Cass, D. 1966. Optimum growth in an aggregative model of capital accumulation: a turnpike theorem. *Econometrica* 34: 833–850.

Cass, D., and K. Shell. 1976. The structure and stability of competitive dynamical systems. *Journal of Economic Theory* 12: 31–70.

Cassel, G. 1918. *Theoretische Sozialökonomie*. Trans. from the 5th German edn. as *Theory of Social Economy*. New York: Harcourt, Brace, 1932.

Coles, J.L. 1985. Equilibrium turnpike theory with constant returns to scale and possibly heterogeneous discount factors. *International Economic Review* 26: 671–680.

Dasgupta, S., and L.W. McKenzie. 1985. A note on comparative statics and dynamics of stationary states. *Economic Letters* 18: 333–338.

de Araujo, A.P., and J.A. Scheinkman. 1977. Smoothness, comparative dynamics, and the turnpike property. *Econometrica* 45: 601–620.

Dechert, W.D., and K. Nishimura. 1983. A complete characterization of optimal growth paths in an aggregated model with a non-concave production function. *Journal of Economic Theory* 31: 332–354.

Dorfman, R., P. Samuelson, and R. Solow. 1958. *Linear programming and economic analysis*. New York: McGraw-Hill.

Epstein, L.G. 1986. Implicitly additive utility and the robustness of turnpike theorems. *Journal of Mathematical Economics* 15: 111–128.

Epstein, L.G. 1987. The global stability of efficient intertemporal allocations. *Econometrica* 55: 329–356.

Evstigneev, I.V. 1974. Optimal stochastic programs and their stimulating prices. In *Mathematical models in economics*, ed. J. Łoś and M. Łoś. Amsterdam: North-Holland.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Gale, D. 1956. The closed linear model of production. In *Linear inequalities and related systems*, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Gale, D. 1967. On optimal development in a multi-sector economy. *Review of Economic Studies* 34: 1–18.

Grandmont, J.M. 1985. On endogenous business cycles. *Econometrica* 53: 995–1046.
Heal, G., and H. Ryder. 1973. An optimum growth model with intertemporally dependent preferences. *Review of Economic Studies* 40: 1–33.
Inada, K. 1964. Some structural characteristics of turnpike theorems. *Review of Economic Studies* 31: 43–58.
Keeler, E.B. 1972. A twisted turnpike. *International Economic Review* 13: 160–166.
Koopmans, T.C. 1960. Stationary ordinal utility and time perspective. *Econometrica* 28: 287–309.
Koopmans, T.C. 1965. The concept of optimal economic growth. In *The econometric approach to development planning*, Pontificae Academiae Scientiarum Scripta Varia No. 28. Amsterdam: North-Holland.
Lucas, R., and N. Stokey. 1984. Optimal growth with many consumers. *Journal of Economic Theory* 32: 139–171.
Magill, M.J.P. 1977. Some new results on the local stability of the process of capital accumulation. *Journal of Economic Theory* 15: 174–210.
Majumdar, M., and T. Mitra. 1982. Intertemporal allocation with a non convex technology: The aggregative framework. *Journal of Economic Theory* 27: 101–136.
Marimon, R. 1984. *General equilibrium and growth under uncertainty: The turnpike property,* Discussion paper No. 624. Evanston: North-western University, August, 1984.
McKenzie, L.W. 1963. Turnpike theorems for a generalized Leontief model. *Econometrica* 31: 165–180.
McKenzie, L.W. 1968. Accumulation programs of maximum utility and the von Neumann facet. In *Value, capital, and growth*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.
McKenzie, L.W. 1974. Turnpike theorems with technology and welfare function variable. In *Mathematical models in economics*, ed. J. Loś and M.W. Loś. New York: American Elsevier.
McKenzie, L.W. 1976. Turnpike theory. *Econometrica* 44: 841–865.
McKenzie, L.W. 1977. A new route to the turnpike. In *Mathematical economics and game theory*, ed. R. Henn and O. Moeschlin. New York: Springer-Verlag.
McKenzie, L.W. 1983. Turnpike theory, discounted utility, and the von Neumann facet. *Journal of Economic Theory* 30: 330–352.
Mill, J.S. 1848. *Principles of political economy.* London: Parker. New edn, London: Longmans, Green, 1909.
Mitra, T. 1979. On optimal growth with variable discount rates: Existence and stability results. *International Economic Review* 20: 133–146.
Negishi, T. 1960. Welfare economics and existence of an equilibrium for a competitive economy. *Metroeconomica* 12: 92–97.
Radner, R. 1961. Paths of economic growth that are optimal with regard only to final states. *Review of Economic Studies* 28: 98–104.
Ramsey, F. 1928. A mathematical theory of savings. *Economic Journal* 38: 543–559.
Samuelson, P.A. 1966. Market mechanisms and maximization. In *The Collected scientific papers of Paul Samuelson*, ed. J. Stiglitz, vol. 1, 425–492. Cambridge, MA: M.I.T. Press.
Samuelson, P.A. 1971. Turnpike theorems even though tastes are intertemporally interdependent. *Western Economic Journal* 9: 21–26.
Samuelson, P.A., and R.W. Solow. 1956. A complete capital model involving heterogeneous capital goods. *Quarterly Journal of Economics* 70: 537–562.
Scheinkman, J. 1976. On optimal steady states of n-sector growth models when utility is discounted. *Journal of Economic Theory* 12: 11–20.
Skiba, A.K. 1978. Optimal growth with a convex-concave production function. *Econometrica* 46: 527–540.
Takahashi, H. 1985. *Characterizations of optimal programs in infinite horizon economies*. PhD thesis, University of Rochester.
Treadway, A.B.. 1971. The rational multivariate flexible accelerator. *Econometrica* 39: 845–855.
von Neumann, J. 1937. Über ein Okonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsätzes. *Ergebnisse Eines Mathematischen Kolloquiums* 8: 73–83. Translated in *Review of Economic Studies* 13, (1945): 1–9.
von Weiszäcker, C.C. 1965. Existence of optimal programs of accumulation for an infinite time horizon. *Review of Economic Studies* 32: 85–104.
Weitzman, W.L. 1973. Duality theory for infinite horizon convex models. *Management Science* 19: 783–789.
Yano, M. 1984a. Competitive equilibria on turnpikes in a McKenzie economy, I: A neighborhood turnpike theorem. *International Economic Review* 25: 695–718.
Yano, M. 1984b. The turnpike of dynamic general equilibrium paths and its insensitivity to initial conditions. *Journal of Mathematical Economics* 13: 235–254.
Yano, M. 1985. Competitive equilibria on turnpikes in a McKenzie economy, II: An asymptotic turnpike theorem. *International Economic Review* 26: 661–670.

Turnpike Theory, a Current Perspective

M. Ali Khan and Adriana Piazza

Abstract

This 2012 perspective of the 1987 *Palgrave* entry on 'turnpike theory' highlights the subsequent development of the subject in the light of a critical re-reading of the original. It distinguishes the 1949 conception, a response of

Samuelson to a 1945 von Neumann challenge to the reception of his growth model in the economic literature, from the more capacious 1976 outline furnished by McKenzie. Thus, it differentiates asymptotic convergence of infinite-horizon optimal programs from what it terms their finite-horizon, classical turnpike counterparts. It identifies a move from the investigation of general theorems to a more detailed working of simple examples, and reports results on specific models of 'choice of technique' in development planning, and of lumber extraction in the economics of forestry. Drawing on ongoing advances in the field of dynamical systems, it sees such models as both litmus tests of the general theory and as productive settings to study the *rationalisability* of policy functions and a 'folk theorem of intertemporal resource allocation'. The entry concludes with brief speculative remarks for future directions.

Keywords

Asymptotic convergence; Classical turnpike theory; Competitive equilibrium; Development planning; Discount factor; Folk theorem; Forest management; Intertemporal resource allocation; Middle-early-late turnpike; Neighborhood turnpike theorem; Optimal programs; Radner's value-loss method; Ramsey model; Rationalisability; Von Neumann growth model

JEL Classifications

C62; D90; Q23

It is common knowledge within the economics profession that the substance underlying the term *turnpike*, and the theorem to which this noun served as an adjective, entered economic theory as a conjecture in a 1949 Rand memorandum authored by Paul Samuelson. Referring to the optimal rate of growth λ^* in the 1935–36 von Neumann growth model, and to the maximal, balanced factor proportions υ^* that underlie it, Samuelson was to write:

> Let us return to the interpretation of the optimal rate of growth λ^*. Growth in the equilibrium mode will ultimately surpass any other rate of balanced growth. This suggests that if we start with any factor proportion $\upsilon \neq \upsilon^*$, it will still pay us if we are investing *for the very far future* to get into (or near) the equilibrium mode. At worst, we can do this by throwing away some of whichever factor is initially redundant as compared to υ^*; and at best we can obviously make some use of the redundant factor. [C]learly, at first υ would not be near υ^* but would ease in gently toward υ^*. And it is also clear that if we prescribe υ_0 and want the maximum υ_n, then as we finally get near n, it will pay to leave υ^* even if we are already there or near there. One would conjecture, therefore, that beginning with $\upsilon_0 \neq \upsilon^*$ and ending with $\upsilon_n \neq \upsilon^*$, the optimal time path of υ would look something like Fig. 14 for large n. As n gets large, the average υ should approach υ^* (S33:489).
>
> [In all references to Samuelson (and Debreu and Koopmans) from (Samuelson 2008) (and (Debreu 1983) and (Koopmans 1970) respectively) the first number indicates the chapter, and the second, the page within it.]

It is also well-known, at least since Lionel McKenzie gave it prominence in his 1976 Fisher–Schultz lecture, that the term itself, as opposed to its underlying substance, occurs in Chapter 12 of a 1958 volume authored by Dorfman et al. (1958) (henceforth DOSSO). One may quote from McKenzie's quotation from DOSSO:

> It is exactly like a turnpike paralleled by a network of minor roads. There is a fastest route between any two points; and if the origin and destination are close together and far from the turnpike, the best route may not touch the turnpike. But if origin and destination are far enough apart, it will always pay to get on to the turnpike and cover distance at the best rate of travel, even if this means adding a little mileage at either end (841).
>
> [All references to McKenzie·s papers are from (McKenzie 1986); this is also the relevant source for papers prior to 1986 not referenced. A number on its own at a quotation·s end refers to the page number in the relevant reference.]

McKenzie was to conclude the first paragraph of his lecture with the statement that 'It is due to this reference, I believe, that theorems on asymptotic properties of efficient, or optimal, paths of capital accumulation came to be known as *turnpike theorems*.'

In terms of a time-line pertaining to *turnpike theory*, the benchmark dates 1960 and 1965–66 follow 1949 and 1958. In a 1960 chapter (S26), an appendix originally written for possible inclusion

in DOSSO, Samuelson returns to the subject under the heading 'efficient paths of capital accumulation in terms of the calculus of variations'. After a reference to the discrete-time treatment in DOSSO, an analogy to the von Neumann growth rate, and a consequent continuous-time conjecture, is adduced:

> For the discrete-time case, the so-called 'turnpike theorem' was enunciated. This asserts that if the goal of maximization is far enough ahead, one will want to travel very near to the von Neumann mode of balanced growth. The economic common sense of this evident, but it would be well to have a general proof for the continuous-time case. The remaining two sections will deal with this and related matters. Economic intuition suggests that the local result proved here must also hold in the large, but this must remain an open question (S26:297).

In addition to its uplifting of maximally balanced growth programs to intertemporally efficient ones, the chapter emphasises two other novel points of view: it (i) reads the theorem as part of the 'formal mathematical analogy between classical thermodynamics and mathematic economic systems' (more fully explored in another contemporaneous chapter (S44)), and it (ii) identifies the catenary property as its local manifestation. The classical Euler–Lagrange first-order conditions of the classical calculus of variations, as in the modernised formulation of Carathéodory and Poincaré, when linearised around the singular point, furnish real eigenvalues that come in 'pairs of opposite signs . . . that lead to the desired catenary motions around the saddle point'. By 1965, the earlier uplifting to four notions of efficiency (a reformalisation, already available in DOSSO, of the 'best route') had been uplifted further to the 1928 Ramsey maximand of an integral of felicities depending on consumption, and with the 'usual 'small vibrations'' analysis of the motions in the neighborhood of the equilibrium point' correspondingly reformulated. However in this, Samuelson (S136; also Fig. 3 and Footnote 8) is in company with Atsumi, Koopmans and Cass, and he will pursue this matter for at least the next five years; see S137–S142, S150, S223–224, K26–K28 (Fig. 10 on pp. 575, 591 (Vol. 1) and Fig. 5, p. 176 (Vol. 2)) and for other references. In any case, by 1966, the original version of the theorem, based on the maximisation of terminal stocks, had also been understood as a global result in a flowering in which Kuhn, Hicks and Morishima in 1961, Radner and Furuya-Inada in 1962, McKenzie in 1963, and Nikaido, Inada and Koopmans in 1964 played an essential part; see S136 (also Footnote 1) and K23 (Footnotes 4–9 and 1–2) for these references.

What this abbreviated narrative omits is that the turnpike conjecture, and the substantial theorems in which it found, and continues to find, its rigorous crystallisation, originated as a local dispute between Samuelson and von Neumann concerning framing: a missed opportunity 'sometime in 1945' for Samuelson to respond to von Neumann's challenge that 'his model of general equilibrium . . . involved new kinds of mathematics which had no relation to the conventional mathematics of physics and maximization' (S130:15); also see S406:75. This is surely not common knowledge within the profession, and despite McKenzie's masterful 1986–1987 surveys, not as well-appreciated even by the *cognoscenti* as it ought; see (McKenzie 1986, 1987). In his 1998 Richard Ely lecture, McKenzie (McKenzie 1998) himself recounts the episode, but takes an altogether rosier view of the outcome than Samuelson himself (a full cigar versus half a cigar), and represents Samuelson's response to von Neumann's claim that 'maximization of an objective function had no part in his theory' as saying that 'maximization would enter once disequilibria were considered'. What was new in von Neumann's 'general equilibrium theory' was of course the notion of a saddle point and the minimax theorem; and as Karlin's 1960 text was to show most transparently, this entailed a separation: the question of the existence of a maximal balanced growth program treated independently of its price characterisation involving a minimal interest rate. Thus, in an ironic twist, the von Neumann use of the fixed-point theorem was not to be brought into play at all in the proof of his theorem; see the polished treatment of the minimax theorem in (Simons 2008, Chapter 1). But Negishi notwithstanding (see (McKenzie 2002)), this separation between existence and characterisation of equilibrium, with fixed-point theory servicing the

first and Hahn–Banach theory the second, and neither concerned with computation, was to leave its footprint on the development of Walrasian general equilibrium theory in the sixth decade of 20th century economics; along with DOSSO and the 1971 Arrow–Hahn (Arrow and Hahn 1971) reworking of Debreu's 1959 classic, see (Khan 2010). Von Neumann's challenge to Samuelson was then to provide an asymptotic implementation of his theorem on 'good' infinite programs using finite-horizon ones that were optimal in the sense conventional at the time. In another, more current, vernacular, this was to ask for a microeconomic foundation, based on optimisation, to the macroeconomic regularities of the maximal–minimal growth–interest pair that von Neumann had uncovered. Such a microfoundation was subsequently supplied by Arrow, Debreu and Scarf (D1, D5, D11 and references) in what can now be seen as the second and third fundamental theorems of welfare economics: to continue using the services of the separating hyperplane theorem to exhibit Pareto optimal and core allocations as Walrasian (price) equilibria, one with redistribution and the other without. But since general competitive analysis was to repress infinite-dimensional commodity spaces, which is to say, repress the 1953–54 analyses of Malinvaud (D5:104) and Debreu (D5), and the 1958 treatments of Hurwicz (Arrow et al. 1958) and Samuelson (S21), and to develop, until 1971 at any rate, in the setting of a finite number of commodities, there was no general equilibrium ground on which turnpike theory could turn and find play. It consequently receded and moved away, von Neumann's contribution bifurcated; however, see (McKenzie 2002, Bewley 2007) as attempts at restitution.

The two arms of this bifurcation were of course a theory of resource allocation premised on heterogeneous agents, a heterogeneity with a cardinality of a finite set or that of a continuum, but with a finite number of commodities; and another premised on a representative agent, or, to coin a phrase, a continuum of representative agents, but with a *bona fide* infinite-dimensional commodity space. And from 1965–66 onwards, turnpike theory was to thrive in the context of the latter. The 1965 Atsumi-von-Weiszäcker reformulation of Ramsey's criterion led rather quickly in 1967–70, at the hands of Gale, McKenzie and Brock, to a complete treatment of the theory of the existence of optimal programs in Ramsey-like situations when the aggregate is not well defined. However, unlike Brock and Gale, McKenzie retained his primary focus on the turnpike theorem, and on its generalisation to a fully multi-sectoral environment. This focus is total, and the first two theorems of his paper surely mark 1968 as an important benchmark date for turnpike theory. As he was to write later in the *Palgrave* (McKenzie 1987):

> The spirit of the original turnpike theorem is not well preserved in the aggregative model since the emphasis in the original theorem lies in the relative composition of the capital stock. Turnpike theorems for the general multi-sectoral model with a Ramsey objective and a von Neumann technology were first proved by Gale (1967) and McKenzie (1968). Their order of proof does not differ from that of Atsumi, which is, in turn, parallel to the proof used by Radner in the model with maximal growth as an objective.

McKenzie sketches the outline of his proof in (McKenzie 1987), and again in his 2002 book (McKenzie 2002). As in (Karlin 1960), one draws on the separating hyperplane theorem to associate a price system to the facet of stationary programs, maximal in the obvious sense, and once this is done, to derive bounds when the finite-horizon optimal program is not 'on or near' the facet of stationary programs. And to be sure, leaving finite-horizon programs aside, and on appealing to the Atsumi-von-Weiszäcker optimality criterion, one also obtains the existence of such infinite-horizon programs. If the stationary program is unique, a sharper existence result is obtainable. In any case, what is clear is that by 1968, *good* and *bad* programs, the *von Neumann facet* and its possible shrinkage to a singleton, as in Brock's seminal 1970 treatment, the emphasis on *bunching* or *clustering* of optimal programs, as opposed to convergence to a unique equilibrium, had all became staples in the vernacular of turnpike theory. [All references in this paragraph are available in (McKenzie 1986; McKenzie 1987).]

The year 1976 takes its place alongside 1949, 1958, 1965–68 as a benchmark year for turnpike theory. In his 1976 Fisher–Schultz lecture, already

referred to above, McKenzie surveys where the matter stood, and observes that 'there were no global results for perhaps the most relevant case for decision making, the maximization of a discounted sum of utility over time with scarce labor, [but that] in the past two years the situation has changed significantly' (843). He cites Scheinkman for a first attempt at what will later become the 'neigborhood turnpike theorem', Rockafellar and Cass and Shell for conditions on the concavity of the felicity function, Brock and Scheinkman for a rendering in continuous-time, Araujo and Scheinkman for a dominant diagonal condition which 'does not translate directly into the degree of concavity or the size of the discount factor', and his own result concerning non-stationary felicities; see (Mitra 2005) for a more recent synthesis. It is important to understand the need for these additional conditions in the discounted case: proofs in the undiscounted case are based heavily on the fact that the value-loss of good programs must converge to zero, thereby guaranteeing their convergence to the von Neumann facet. This is no longer true in the discounted case, as also brought out in (McKenzie 1987, p. 715). But there is another less positive reason for singling out 1976. It is the year when turnpike theory, deprived of its moorings in general equilibrium theory, over-reaches. McKenzie puts forward a tri-partite classification of turnpike theory – the middle, the early and the late turnpike – and with this classification, turns a local challenge regarding asymptotic implementation of an infinite-horizon program into the global study of intertemporal allocation of resources. More specifically, he widens the meaning of the word *asymptotic* to include, under the category of the late turnpike, the convergence of optimal infinite-horizon programs to stationary programs, and to their tendency to 'bunch or cluster' together in the far future. Indeed, the references cited in his 1976 lectures all pertain to the *late* turnpike. Thus, in his *Handbook* chapter a decade later, McKenzie was to write:

> The theory ... will cover both discounted and undiscounted utility. We will seek to determine the asymptotic behavior of maximal paths, which display a tendency to cluster in the sufficiently distant future from whatever capital stocks they start. In models with stationary utility functions the clustering has been seen as convergence to a stationary path along which capital stocks are constant. The existence of infinite optimal paths in the stationary disaggregated model were proved by Gale (1967). Asymptotic theorems in this model were proved by Atsumi (1965), Gale (1967), and McKenzie (1968), ... [and they] were extended to [discounted] models by Scheinkman (1976) and by Cass and Shell (1976). Excellent examples from the theory of competitive equilibrium are the recent works of Becker (1980), Bewley (1982) and Yano (1981), where the turnpike results ... are used to prove that competitive equilibria approach stationary states over time. It has been suggested that our subject is best described as the study of economizing over time.

This extended quotation, with its going over of references already read and referenced, is necessary to bring out how this widening of the term *asymptotic*, and its heightened elevation to the entirety of the qualitative theory of economic dynamics, robs turnpike theory of its very identity. The *late* turnpike is not a turnpike, in the sense that it typically requires infinite time to get on it, and with no terminal capital stock there is 'no getting off it'. In the words of Koopmans (K:201, volume 2), 'there is no fatal cut-off point', unlike the early and middle turnpike twins. In the quotation above, other than those to Atsumi and McKenzie himself, all other references pertain to the late turnpike, rather than to the *early* and *middle* ones, and by 1998, the terminology proliferates to make Ramsey, von Neumann and Samuelson serve as adjectives to the noun 'turnpike'; see (McKenzie 1998, Section II and III). A sustained case for the rescue and protection of the original Samuelsonian conception of the theory from its more universalistic 1976 ambitions was first made by Khan and Zaslavski (2010) in 2010.

McKenzie's 1976 reading was immensely influential: it did not simply balance work on the early and middle turnpike by those on the late turnpike, but (barring a few exceptions) entirely eliminated the former from the mainstream journals, handbooks and anthologies; see (Benhabib 1992; Majumdar et al. 2000; Aghion and Durlauf 2005; Dana et al. 2006). Thus, in Mitra's important 2005 synthesis of the 1976 results, it is 'clarified that it is only in the sense of 'global asymptotic stability"

that the term 'turnpike property" is used in this paper'; also see (Khan and Piazza 2011) for such a trajectory extending from the 1977 Araujo and Scheinkman analysis to Bewley's 2007 text. Indeed, McKenzie begins his own *Palgrave* entry with the classical and neoclassical economists, as represented by Mill and Cassel, and originates the subject in the time-honoured conception of the 'eventual convergence of the economy to a stationary state as a consequence of the growth of population and the accumulation of capital, in the absence of continual technical progress or continual expansion of natural resources'. He opens the penultimate paragraph of his entry with the summary statement 'The theorems that have been reviewed are all concerned with the convergence of optimal paths to stationary optimal paths' – a statement that applies even more so to the references in the concluding paragraph. In what is surely a classic, divided into five sections – turnpike theorems for the von Neumann model, for the Ramsey model, for Ramsey models with discounting, for competitive equilibria, and for generalisations to habit formation uncertainty and non-convex technologies – only the first concerns the early and middle turnpike, and the hazard of an opening getting reified into a doctrinal closing was unfortunately realised; see (Khan and Piazza 2011), and especially their introduction, for a substantiation of this reading. The point of course is that it is the classical conception that was novel to both economic theory and to applied mathematics, and it was not a re-packaging of the qualitative theory of differential and difference equations, albeit ones generated by the Euler–Lagrange conditions of the calculus of variations or their discrete-time counterparts. Indeed, even though Radner's value-loss methods play a role in the important work of Araujo and Scheinkman on the asymptotic convergence of optimal infinite-horizon trajectories, it is the implicit function and Hirsch–Pugh theorems, time-tested stalwarts of dynamical-systems theory, that bear the brunt of the lifting; (Carlson et al. 1991, Zaslavski 2005) may be the only mathematical writers to make and appreciate the distinction.

With this 2012 re-reading of a 1987 reading behind us, we can turn to subsequent developments and ask whether there is very much to say, if anything, about turnpike theory in the classic Samuelsonian sense. However, prior to this, we need to highlight 1986 as another benchmark date for turnpike theory. It is in that year that Boldrin, Deneckere, Montrucchio and Pelikan (henceforth BDMP) investigate the relevance of the 1972–1974 Sonnenschein–Mantel–Debreu theorems (D16 with its five references) for the theory of optimal intertemporal allocation of resources; see (Boldrin and Montrucchio 1986a, Boldrin and Montrucchio 1986b, Deneckere and Pelikan 1986) and reference to antecedent work by Montrucchio (1984). They show that any twice continuously differentiable function can be *rationalised* as the policy function of an appropriately defined dynamic optimisation model, and taking their cue from May (1976) and the twice-differentiability of the logistic map, they conclude that anything, including chaotic trajectories, can emerge as a result of optimisation by a representative agent over infinite time. The implications of this work for turnpike theory (middle, early or late) could hardly be minimised; see (Dana et al. 2006, Chapters 4 and 6) and (Khan and Piazza 2011) for an outline of the theory as it has taken shape since 1986, and for its reliance on Sharkovsky's theorem. However, the point is that this work has different implications depending on which of the two conceptions of the subject is presupposed. If one is focused on the late turnpike, and on the asymptotic convergence of optimal trajectories to the benchmarks of a stationary or a quasi-stationary model (the terms are McKenzie's (McKenzie 1986)), it surely delivers a resounding, if not fatal, refutation to the turnpike conjecture. If, on the other hand, one is focused on the early or middle turnpikes, and on the approximation of infinite- by finite-horizon programs with a specified terminal capital stock, the turnpike conjecture remains very much alive and viable. If confronted with an optimally chaotic solution to infinite-horizon problem, it can still ask whether an optimal solution to a finite-horizon problem, for a large enough horizon reflecting the tolerable error of approximation, stays close to the infinite one – partakes and participates in the chaos of its parent, so to speak.

Samuelson's 1976 periodic turnpike theorem (S224) may already be a pioneering answer to a question yet to be formulated and asked, the only (not inessential) difference being that in his case the periodicity is exogenous rather than endogenous. Indeed, putting the question this way then leads to giving a constructive twist to what may have so far seemed to be a purely semantic issue, a rather negative attempt to distinguish, and dissociate, Samuelson from McKenzie. This is then to ask whether there is a relationship between their two differing conceptions of turnpike theory: the local versus the global – local in the sense of capital theory being a locality of the general environs of intertemporal resource allocation. Put more sharply, can a result on the asymptotic convergence deliver a classical turnpike theorem as its corollary?; and conversely, can a classical turnpike theorem imply, in the limit, as the time-horizon is extended without bound, a coming-together of the optimal trajectories? For initial exploratory studies, see (Khan and Zaslavski 2010), and subsequent to that work, (Khan and Piazza 2011; Zaslavski 2009; Zaslavski 2010).

However, even before the BDMP floodgates opened, Scheinkman's 1976 question as to the extent to which global asymptotic stability of optimal infinite-horizon programs could be salvaged from an undiscounted to a discounted setting, hovered over the subject. In 1983, McKenzie writes:

> Asymptotic theory for optimal paths of capital accumulation is more difficult when the utility function for the single period is concave, but not strictly concave. However, in the case of stationary models where future utility is not discounted, the theory is rather fully developed in McKenzie (1968, 1976). In the case of discounted utility and quasi-stationary models this order of proof does not succeed, because convergence of optimal paths to the facets on which optimal stationary paths lie cannot be proved to be asymptotic on the basis of arguments from value losses, or utility gains. In order to carry the argument further, we must use the convergence of the von Neumann facets associated with discount factors to the von Neumann facet of the undiscounted model as the discount factor approaches unity (330–331).
>
> [Under the adopted bibliographic convention, McKenzie (1983) is available in (McKenzie 1987).]

This is the genesis of what has come to be called McKenzie's *neighborhood turnpike* theorem: 'the larger the neighborhood chosen, the smaller the discount factor allowed', The theorem was answered, potentially negatively, but more than a decade later. In 1995, Nishimura and Yano were to show that for any discount factor arbitrarily close to unity, one could construct a two-sector LS (Leontief–Shinkai) model whose optimal policy function would exhibit ergodically chaotic (not simply topologically chaotic) optimal trajectories. The question of course is how compelling this answer was: McKenzie speaks of a *given* model and Nishimura and Yano respond in terms of a *constructed* model. In any case, the following year the profession was to be presented with even more remarkable results: Mitra and Nishimura and Yano were to discover, entirely independently, 'exact discount factor restrictions for dynamic optimization models'. Loosely speaking, this was to identify, in the context of a general model, a threshold for the discount factor that was necessary and sufficient for an optimal trajectory in that model to exhibit periodthree cycles, and therefore by Sharkovsky's theorem, cycles of all periods; for details and references, see (Majumdar et al. 2000, Chapters 11, 12) and (Dana et al. 2006, Chapter 4). It was then only natural that within a decade of these achievements, the three introductory paragraphs of McKenzie (1983) would be consummated in a precise 'folk-theorem of optimal accumulation'. The theorem was to announce, in the context of a general model, the potential existence of a threshold discount factor, such that optimal trajectories would exhibit chaotic behaviour for all factors less than that threshold, and asymptotic convergence, which is to say, a late turnpike for those above it; see (Khan 2005) for an initial formulation. And not unlike its game-theoretic cousin, this folk theorem, even in its precise formulation, still remains a conjecture for a general-enough class of models; see (Khan and Mitra 2010; Khan and Mitra 2011; Khan and Mitra 2012).

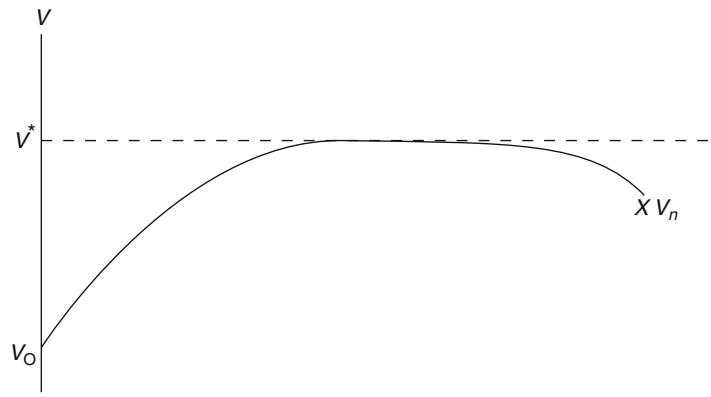
The BDMP work, and subsequent papers that gave it grounding, had, by necessity, considered simple and specific, typically two-sector, production structures. As an unintended consequence,

this then changed the tone of the subject: it retreated from its aspiration towards increasing generality – in addition to non-convexity and uncertainty, even an embrace of non-stationarity – to one looking towards the explicit and detailed working of specific examples. In this, it was also spurred on by the theory of dynamical systems, and more specifically, 'iterated functional systems (IFS)'; see Barnsley (2006) and his references, and for their influence on economic theory, see (Bhattacharya and Majumdar 2004, 2007) for references to papers of Bhattacharya, Majumdar, Mitra and others. In particular, the aggregative model popularised by Ramsey, and Solow in 1956 (currently the RCK [Ramsey–Cass–Koopmans] workhorse in macroeconomics) and the two-sector versions swirling around the LS (Leontief–Shinkai) model already mentioned in connection with the work of Nishimura and Yano, and investigated further in Fujio (2009), were to be complemented by two settings originating in subjects not apparently related to each other, and both neglected only until recently. The first dates to the early 1960s, when it served as the lightning rod for polemics between two transatlantic locations, and was dusted off and rediscovered as the so-called RSS (Robinson–Solow–Srinivasan) model by Khan and Mitra (2005); see (Khan and Mitra 2005) for details as to genealogy and references. The second is also a re-visiting, but this time of Samuelson's work on the economics of forestry (S218) as the MW (Mitra–Wan) model; see (Khan 2005, Khan and Piazza 2012) for references to the pioneering papers. Both involve a period-by-period allocation of an inelastically supplied resource: in the first case, labour to produce one or more types of machine chosen from a finite set, or a consumption good along with any of the machines that may be available; and in the second case, land partitioned out among a finite set of tree-vintages, all costlessly grown. Machines depreciate at the same rate, trees grow and yield lumber at different rates. The objective is to maximise an undiscounted, or discounted, stream of the consumption good, with or without strictly concave, or even concave, felicities; see (Khan and Zaslavski 2009; Khan and Piazza 2011; Khan and Piazza 2012). Thus both settings are multi-sectoral, and thereby immune to McKenzie's complaint regarding the RCK model as not in keeping with the original spirit of turnpike theory. The point, however, is that, simple as these two settings are, they are not simple enough! On the one hand they have led to a transparent geometric consolidation, and considerable sharpening, of the theory, but on the other hand, the dynamics they exhibit are rich enough to defy full understanding even for special cases of a single type of machine or a dual-aged forest. Work is ongoing, but has already yielded insights for capital theory in the large that go beyond the two instances themselves; for existence and asymptotic convergence, rationalisability, chaotic dynamics, parametric restrictions, policy correspondences and several bifurcations revealing an unexpected intricacy to the 'folk theorem', see respectively (Khan and Zaslavski 2009; Khan and Piazza 2010; Khan and Piazza 2010; Khan and Piazza 2012; Khan and Piazza 2011; Khan and Mitra 2012; Khan and Mitra 2011; Khan and Mitra 2012; Khan and Mitra 2010) and their references.

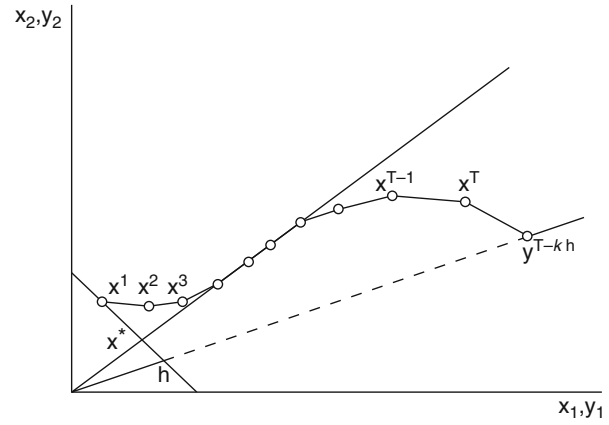
The above two paragraphs concern asymptotic convergence and the late turnpike: the question concerning progress on the classical turnpike theorem, asked above and left hanging, hangs still. It is of interest that it is the early work on non-classical environments, non-convex technologies and uncertainty, that scrupulously maintained a distinction between the two conceptions and preserved the autonomy of the original. The 1982 analysis of Majumdar and Nermuth (1982) in the one case, and the 1998 analysis of Joshi (1998), building on the pioneering 1978 connection to the martingale property in Follmer and Majumdar (1978), on the other, is surely worth careful study even today; also see (Bhattacharya and Majumdar 2007; Arkin and Evstigneev 1987) and their references. More recently, even in a deterministic, non-stochastic context, there has been a breaking of new ground in the context of the RSS and MW models delineated above. This concerns a further weakening of the optimality notion. Thus, rather than the Samuelsonian triple limit (S130:15), an interesting 'quarter limit' seems to be involved in the four separate

considerations that are quantified. In other words, for any given levels of the initial and terminal capital stocks, υ_0 and υ_1, and for any three given levels of approximation, ε_1, ε_2 and ε_3, one can find a large but finite time horizon $T(\upsilon_0, \upsilon_1, \varepsilon_1, \varepsilon_2, \varepsilon_3)$ such that any ε_3-optimal program starting from υ_0 and required to furnish υ_1 at its termination, spends $(1 - \varepsilon_1)$-proportion of the time in an ε_2-proximity to the turnpike for all time periods that extend over *T*. Such results were proved for the RSS model in (Khan and Zaslavski 2010), and for the MW forestry model in (Khan and Piazza 2011). In each case, the theorems are shown to yield *uniform* asymptotic convergence of the maximal stationary programs and thereby generalise corresponding results in (Khan and Zaslavski 2006; Khan and Piazza 2010). With so many epsilons, surely *nonstandard analysis* waits in the wings; it may also be the right mathematical language to express McKenzie's notions of 'bunching and clustering' in the context of the classical Samuelsonian conjecture, and make the first steps towards a set-valued turnpike theory. To return to a theme broached earlier, the exploratory analysis presented in (Khan and Piazza 2011) bunches the full continuum of periodic programs together, and investigates the entire set as a possible candidate for a turnpike. Thus, instead of a freeway or turnpike, the relevant metaphor here is that of an air- or sea-lane through which journeys are routed, even though those lanes may not be the most direct route. Within the lane there are many possible routes, and which particular route is taken on one occasion is not the most relevant consideration for another. From a technical point of view, we then substitute a set for a point to obtain a non-trivial generalisation of the theory that reduces to the standard one when the set shrinks to a point and the sufficient conditions of the result are automatically activated.

For a field of inquiry to live, it must furnish problems, substantive and technical, to be worked on, and must have a space for its advances to be categorised and catalogued. The clearing of terminological confusion is important only in so far as it advances this end. Whether an economy, perfectly or imperfectly competitive, tends in the long run to a stationary state, is a time-honoured and classical question that surely dates to the 19th century, if not earlier. The turnpike conjecture of the mid- 20th century is different. It asks whether an allocation of resources, efficient from the viewpoint of one generation, however measured, must be near a stationary state, or some other benchmark which is generation-independent, given what each generation has received, and given what it has perforce to leave. Put another way, and again leaving aside a precise definition of how long a generation is, its theorems answer how generational well-being is to be ensured under the constraint that one bequeath to the future the world no worse than the one received. It is now recognised that our own century is beset with a whole host of multi-faceted environmental difficulties, and given that population and institutional design are once again being conceived as manipulable instruments of policy, turnpike theory surely has a bright future in it; see S220–S222,



Turnpike Theory, a Current Perspective, Fig. 1 Figure 14 in Samuelson (1949)



Turnpike Theory, a Current Perspective, Fig. 2 Figure IX in Koopmans (1964)

S234–S236, K6–K7, K9–K14 in volume 2 and the relevant essays in (Aghion and Durlauf 2005) for hints. In any case, in an age of computation and experimental mathematics (Barnsley 2006; Borwein and Bailey 2008), asymptotic convergence, and the *long run* that it implies, needs approximation to be rendered operational and policy-relevant, thereby guaranteeing an inevitable slide of the first question into the second, even for conventional fields of inquiry as in Yano (1990, 1998) or in McKenzie's (1998) singling out of new growth theory. In conclusion then, what began as a 25-year re-reading and updating of a 1987 entry, has, by necessity, had to read a narrative now over 60 years old, a reading that brings out its substantive and technical relevance involving computation for a variety of issues; capital-theoretic to be sure, but broadened enough to pertain to both intertemporal allocation and preservation of resources, including climate.

Acknowledgments

The authors are grateful to Professor Steven Durlauf for seeing the need for an updating of McKenzie's 1987 *Palgrave* entry, and for inviting them to attempt it. In addition to him, the authors thank Professors Sumit Joshi, Maia Linask, Mukul Majumdar, Toru Maruyama, Lionel McKenzie, Tapan Mitra, Boris Mordukhovich, Roy Radner, Paul Samuelson and Sasha Zaslavski for, as the case may be, discussion, encouragement and collaboration over the years. However, errors and idiosyncracy of interpretation are solely the authors'. Adriana Piazza gratefully acknowledges the financial support of Project ANILLO ACT-88 and that of FONDECYT under project 11090254.

See Also

- ▶ Ramsey Model
- ▶ Turnpike Theory
- ▶ Von Neumann Ray

Bibliography

Aghion, P. and Durlauf, S. N. (eds.) 2005. *Handbook of economic growth*. Vols. 1 and 2. Amsterdam: North-Holland.

Arkin, V., and I. Evstigneev. 1987. *Stochastic models of control and economic dynamics*. New York: Academic.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden Day.

Arrow, K. J., Hurwicz, L. and Uzawa, H. (eds.) 1958. *Studies in linear and non linear programming*. Palo Alto: Stanford University Press.

Barnsley, M.F. 2006. *Superfractals*. Cambridge, MA: Cambridge University Press.

Benhabib, J. 1992. *Cycles and chaos in economic equilibrium*. Princeton: Princeton University Press.

Bewley, T.F. 2007. *General equilibrium, overlapping generations models and optimal growth theory*. Cambridge, MA: Harvard University Press.

Bhattacharya, R., and M. Majumdar. 2004. Dynamical systems subject to random shocks: An introduction. *Economic Theory* 23: 1–12.

Bhattacharya, R. and Majumdar, M. (eds.) 2007. *Random dynamical systems*. Cambridge, MA: Cambridge University Press.

Boldrin, M., and L. Montrucchio. 1986a. Cyclic and chaotic behavior in intertemporal optimization models. *Mathematical Modelling* 8: 697–700.

Boldrin, M., and L. Montrucchio. 1986b. On the indeterminacy of capital accumulation paths. *Journal of Economic Theory* 40: 26–39.

Borwein, J., and D. Bailey. 2008. *Mathematics by experiment*. 2nd ed. Wellesley: A. K. Peters.

Carlson, D.A., A.B. Haurie, and A. Leizarowitz. 1991. *Infinite horizon optimal control: Deterministic and stochastic systems*. Berlin: Springer.

Dana, R. A., Le Van, C., Mitra, T. and Nishimura, K. (eds.) 2006. *Handbook of optimal growth*. Vol. 1. Berlin: Springer.

Debreu, G. 1983. *Mathematical economics*. Cambridge, MA: Cambridge University Press.

Deneckere, R., and S. Pelikan. 1986. Competitive chaos. *Journal of Economic Theory* 40: 13–25.

Dorfman, R., R.M. Solow, and P.A. Samuelson. 1958. *Linear programming and economic analysis*. New York: McGraw-Hill Book Co..

Follmer, H., and M. Majumdar. 1978. On the asymptotic behavior of stochastic economic processes. *Journal of Mathematical Economics* 5: 275–287.

Fujio, M. 2009. Optimal transition dynamics in the Leontief two-sector growth model with durable capital: The case of capital intensive consumption goods. *Japanese Economic Review* 60: 490–511.

Gale, D. 1970. Nonlinear duality and qualitative properties of optimal growth. In *Integer and nonlinear programming*, ed. J. Abadie, 309–319. Amsterdam: North-Holland.

Joshi, S. 1998. Turnpike theorems in nonconvex nonstationary environments. *International Economic Review* 28: 225–248.

Karlin, S.J. 1960. *Mathematical methods and theory in games, programming and economics*. Reading MA: Addison-Wesley.

Khan, M. Ali. 2005. Intertemporal ethics, modern capital theory and the economics of forestry, Chapter 2. In *Sustainability, economics and natural resources: Economics of sustainable forest management*, ed. S. Kant and A. Berry, 39–65. Netherlands: Springer.

Khan, M.Ali. 2010. La concorrenza perfetta come teoria dell'equilibrio. In *La matematica*, ed. C. Bartocci and P. Odifreddi, Vol. IV. Rome: Einaudi (English translation available as *Perfect Competition as Equilibrium Theory*).

Khan, M. Ali, and T. Mitra. 2005. On choice of technique in the Robinson–Solow–Srinivasan model. *International Journal of Economic Theory* 1: 83–109.

Khan, M. Ali, and T. Mitra. 2010. *Discounted optimal growth in the two-sector RSS model: A further geometric investigation*. mimeo: Johns Hopkins University.

Khan, M. Ali, and T. Mitra. 2011. *Complicated dynamics and parametric restrictions in the Robinson-Solow-Srinivasan model*. mimeo: Cornell University.

Khan, M. Ali, and T. Mitra. 2012a. Long run optimal behavior in a two-sector Robinson–Solow–Srinivasan model. *Macroeconomic Dynamics* 16: 70–85.

Khan, M. Ali, and T. Mitra. 2012b. Impatience and dynamic optimal behavior: A bifurcation analysis of the Robinson–Solow–Srinivasan model. *Nonlinear Analysis* 75: 1400–1418.

Khan, M. Ali, and A. Piazza. 2010a. On uniform convergence of undiscounted optimal programs in the Mitra–Wan forestry model: The strictly concave case. *International Journal of Economic Theory* 6: 57–76.

Khan, M. Ali, and A. Piazza. 2010b. On the non-existence of optimal programs in the Robinson–Solow–Srinivasan (RSS) model. *Economics Letters* 109: 94–98.

Khan, M. Ali, and A. Piazza. 2011a. Classical turnpike theory and the economics of forestry. *Journal of Behavioral Economics and Organization* 79: 194–201.

Khan, M. Ali, and A. Piazza. 2011b. The concavity assumption on felicities and asymptotic dynamics in the RSS model. *Set-Valued and Variational Analysis* 19: 135–156.

Khan, M. Ali, and A. Piazza. 2011c. Optimal cyclicity and chaos in the 2-sector RSS model: An anything-goes construction. *Journal of Economic Behavior and Organization* 80: 397–417.

Khan, M. Ali, and A. Piazza. 2011d. An overview of turnpike theory: Towards the discounted deterministic case. *Advances in Mathematical Economics* 14: 39–67.

Khan, M. Ali, and A. Piazza. 2011e. The economics of forestry and a set-valued turnpike of the classical type. *Nonlinear Analysis* 74: 171–181.

Khan, M. Ali, and A. Piazza. 2012. On the Mitra–Wan forestry model: A unified analysis. *Journal of Economic Theory* 147: 230–260.

Khan, M. Ali, and A.J. Zaslavski. 2006. On a uniform turnpike of the third kind in the Robinson–-Solow–Srinivasan model. *Journal of Economics* 92: 137–166.

Khan, M. Ali, and A.J. Zaslavski. 2009. On existence of weakly maximal programs: The RSS model with non-concave felicities. *Journal of Mathematical Economics* 45: 624–633.

Khan, M. Ali, and A.J. Zaslavski. 2010. On two classical turnpike results for the Robinson–Solow–Srinivasan (RSS) model. *Advances in Mathematical Economics* 13: 47–97.

Koopmans, T. C. 1970. *Scientific papers of Tjalling C. Koopmans*. Vols. 1 and 2. Berlin: Springer.

Majumdar, M., and M. Nermuth. 1982. Dynamic optimization in non-convex models with irreversible investment: Monotonicity and turnpike results. *Zeitschrift für Nationalökonomie* 42: 339–362.

Majumdar, M. Mitra, T. and Nishimura, K. (eds.) 2000. *Optimization and chaos*. Berlin: Springer.

May, R.B. 1976. Simple mathematical models with very complicated dynamics. *Nature* 40: 459–467.
McKenzie, L.W. 1986. Optimal economic growth, turnpike theorems and comparative dynamics. In *Handbook of mathematical economics*, ed. K.J. Arrow and M. Intrilligator, Vol. 3, 1281–1355. New York: North-Holland.
McKenzie, L.W. 1987. Turnpike theory. In *The new palgrave*, ed. J. Eatwell, M. Milgate, and P.K. Newman. London: MacMillan.
McKenzie, L.W. 1998. Turnpikes. *American Economic Review, Papers and Proceedings* 88: 1–14.
McKenzie, L.W. 2002. *Classical general equilibrium theory*. Cambridge, MA: The MIT Press.
Mitra, T. 2005. Characterization of the turnpike property of optimal paths in the aggregative model of intertemporal allocation. *International Journal of Economic Theory* 1: 247–275.
Samuelson, P.A. 1965–2011. *The collected scientific papers of Paul A. Samuelson*. Vol. 1–7. Cambridge, MA: MIT Press.
Simons, S. 2008. *From Hahn–Banach to monotonicity*. Berlin: Springer.
Yano, M. 1990. von Neumann facets and the dynamic stability of perfect foresight equilibrium paths in neo-classical trade models. *Journal of Economics* 51: 27–96.
Yano, M. 1998. On the dual stability of a von Neumann facet and the inefficacy of temporary fiscal policy. *Econometrica* 66: 427–451.
Zaslavski, A.J. 2005. *Turnpike properties in the calculus of variations and optimal control*. New York: Springer.
Zaslavski, A.J. 2009. Two turnpike results for discrete-time optimal control systems. *Nonlinear Analysis* 71: 902–909.
Zaslavski, A.J. 2010. Structure of approximate solutions for discrete-time optimal control systems arising in economic dynamics. *Nonlinear Analysis* 73: 952–970.

Turnpike Trusts

Dan Bogart

Abstract

Turnpike trusts were private organisations that built and operated toll roads in Britain and the United States during the 18th and 19th centuries. They emerged in 17th century Britain because local governments were unwilling to invest in roads. They issued bonds to finance investment and imposed tolls on road users. In Britain, travel times and freight charges declined by over 40% during 1750–1800; they fell also in the United States. Turnpike trusts also raised land values and promoted urbanisation. They show how changes in institutional arrangements can encourage infrastructure investment and promote economic development.

Keywords

Land values; Turnpike bonds; Turnpike trusts; Urbanization

JEL Classifications

N4

Turnpike trusts were widely employed private organizations that built and operated toll roads in Britain and the United States during the 18th and 19th centuries.

Britain and the United States relied heavily on road transport during their early stages of economic development. They faced a problem, however, because their existing road network was ill-suited for the rising volume of traffic, in particular the growing use of large wagons and carriages. In both economies the demand for road improvements was ultimately satisfied through an institutional innovation known as the 'turnpike trust' or the 'turnpike company'.

Britain had a large network of roads and pathways as early as the 16th century. Although the network was called the 'Kings Highway', responsibility for maintenance was placed upon local governments known as parishes. Parishes financed road improvements by forcing their residents to work without pay and by levying property taxes. This method of financing was satisfactory in a pre-industrial economy, in which road improvement costs were low and traffic was largely internal to the parish. Conditions changed during the 17th and 18th centuries, when wages increased and inter-regional trade and travel began to grow. Each of these factors contributed to a divergence between the road expenditure that parishes were willing to provide and the amount needed for an improved network.

Turnpike trusts emerged as a solution to this problem. Trusts were promoted by landowners and merchants, who lobbied for an act of parliament. Each act transferred authority from parishes to a body of trustees composed of the promoters and other local property owners. Trustees were given the right to finance improvements along a particular road by levying tolls and issuing bonds. The tolls were efficacious because they forced road-users to contribute to the costs of improvement, whereas the bonds helped to mobilise funds for initial investment. The act also placed a number of restrictions on trustees. For example, they could not charge tolls above a maximum schedule, and they could not earn direct profits. Instead, it was expected that trustees would benefit indirectly through higher property values (Albert 1972).

The first turnpike act was passed in 1663, and applied to a short section of the Great North Road connecting London with Leeds, York, and Newcastle. The second turnpike act was not passed until 1695, and it was not until the 1720s that trusts became common along the major highways leading into London. Between 1750 and 1770 turnpike trusts diffused throughout much of the road network, especially in the industrialising areas of the West Midlands and the North. After 1770, the network continued to expand, even as canals were being built. By 1840 there were around 1,000 turnpike trusts managing 20,000 miles (Pawson 1977).

The British colonies of North America inherited the original system, in which roads were free and local governments – parishes, towns, or counties – were responsible for maintenance and improvement. A similar problem emerged where local governments were unwilling to pay for road improvements, despite an increasing need for investment. Significant institutional changes did not occur until after the American Revolution, when states began passing legislative acts creating turnpike companies (Durrenberger 1931).

US turnpikes companies were similar to British turnpike trusts, except they were corporations and financed most of their investments by issuing stock. Turnpike companies were widely established in New England and the Middle Atlantic states between 1792 and 1845. The early companies were adopted along roads linking major cities such as Boston, Philadelphia, and New York with smaller cities in their western hinterlands. The later turnpike companies generally built and operated roads that led to other turnpikes and canals. By 1845 there were over 800 turnpike companies managing approximately 15,000 miles (Klein and Majewski 2004).

In Britain and the US the official rationale for creating turnpikes was that the 'ordinary' laws for repairing highways needed to be amended if the roads were to be improved. Did turnpike trusts and turnpike companies meet these expectations? In Britain turnpike trusts were generally successful in increasing road maintenance and investment. On average, they spent between 10 and 20 times more than the parishes they replaced. Most trusts purchased land and materials in order to widen their roads and improve the surface. Many trusts also spent substantial sums on maintenance, as they had to deal with the growing volume of traffic in the 18th century (Bogart 2005a). US turnpike companies were most successful in raising road investment. The amount of capital raised through stock issues was particular striking given that dividends were rarely paid (Klein 1990). This contrasts with British turnpike bonds, which usually yielded a return of between 4% and 5% (Albert, 1972). American turnpike companies had more difficulties paying for maintenance, in part because traffic volumes were lower, but also because companies had difficulties collecting tolls (Klein 1990).

How were road-users affected by the rise of turnpike trusts and turnpike companies? The evidence for British transport costs shows that the benefits from improved roads substantially outweighed the burden of the tolls, as travel times and freight charges declined by over 40% between 1750 and 1800 (Pawson 1977; Bogart 2005b). The evidence from the United States suggests a similar pattern, in which travel times and freight charges fell after turnpike companies improved the road (Durrenberger 1931). The accounts of contemporaries also suggest that turnpike trusts raised land values and contributed to urbanization. These indirect benefits were especially important because they provided added motivation for landowners

and merchants to promote turnpikes and purchase their stocks and bonds.

Turnpikes are often viewed alongside the canal companies and railroads that superseded them in the second quarter of the 19th century. Improving a road was far less expensive, and therefore the turnpike movement did not lead to domestic and international capital flows as with canals and railways. The benefits of turnpikes were also smaller given the natural limits of horse-drawn transport. That said, one should recognise that turnpikes generated substantial benefits in their era (Pawson 1977; Bogart 2005b). At a time when local and central governments were largely ineffective, these organizations provided a mechanism by which transport investment could be implemented. Their history also provides an illustration of how changes in institutional arrangements can encourage infrastructure investment and promote economic development.

See Also

- ▶ Growth and Institutions
- ▶ Industrial Revolution
- ▶ Tragedy of the Commons

Bibliography

Albert, W. 1972. *The turnpike road system in England 1663–1840*. Cambridge: Cambridge University Press.

Bogart, D. 2005a. Did turnpike trusts increase transportation investment in eighteenth-century England? *Journal of Economic History* 65: 439–68.

Bogart, D. 2005b. Turnpike trusts and the transportation revolution in eighteenth century England. *Explorations in Economic History* 42: 479–508.

Durrenberger, J. 1931. *Turnpikes: A study of the toll road movement in the Middle Atlantic States and Maryland*, 1968. Cos Cob: John E. Edwards.

Klein, D. 1990. The voluntary provision of public goods? The turnpike companies of early America. *Economic Inquiry* 28: 788–812.

Klein, D., and Majewski, J. 2004. Turnpikes and toll roads in nineteenth century America. In *EH. Net Encyclopedia*, ed. R. Whaples. Online. Available at http://eh.net/encyclopeida/article/Klein.Majewski.Turnpikes. Consulted 3 Nov 2005.

Pawson, E. 1977. *Transport and economy: The turnpike roads of eighteenth century Britain*. New York: Academic.

Tversky, Amos (1937–1996)

Eldar Shafir

Abstract

Amos Tversky (1937–1996), a cognitive psychologist, is regarded as a giant in the study of human judgment and decision making, and one of the founders of behavioural economics. His early work in mathematical psychology focused on choice, similarity and measurement. With Daniel Kahneman, he collaborated on a highly influential study of judgmental heuristics and biases, and later published a seminal paper on prospect theory, a descriptive theory of individual choice. These projects have had a revolutionary impact on the study of judgment and decision making. Tversky's work has been influential across many disciplines; he won many awards for diverse accomplishments.

Keywords

Behavioural economics; Decision making; Intransitive preferences; Judgment; Kahneman, D.; Mathematical psychology; Measurement; Normative theory; Preference reversals; Prospect theory; Risk aversion; Similarity models; Tversky, A.; Utility maximization; Risk seeking; Endowment effects; Support theory; Probability judgments; Econometric society

JEL classifications

B31

Amos Tversky, a cognitive psychologist, is regarded as a giant in the study of human judgment and decision making, and one of the founders of behavioural economics. Born on 16 March 1937 in Haifa, Israel, his father was a veterinarian and his mother was a social worker and member of the first Israeli Parliament and those following, for some 15 years. Tversky received his BA from

the Hebrew University in Jerusalem in 1961, majoring in philosophy and psychology, and a Ph.D. in psychology from the University of Michigan in 1965.

The Early Work

Tversky's early work in mathematical psychology focused on the study of individual choice behaviour and the analysis of psychological measurement, exploring almost from the beginning the surprising implications of simple and intuitively compelling psychological assumptions. In one early work, Tversky (1969) showed how a series of pair-wise choices could yield intransitive patterns of preference. To do this, he created a set of options such that differences on an important dimension were negligible between adjacent alternatives, but proved to be consequential once compounded across a number of options, yielding a reversal of preference between the first and the last. This pattern not only contradicted a fundamental assumption of utility theory; it also provided a revealing glimpse into the psychological processes involved in decisions of this kind.

For another example, Tversky's (1977) highly influential model of similarity made a number of simple psychological assumptions: items are mentally represented as collections of features, with the similarity between items an increasing function of the features that they have in common, and a decreasing function of their distinct features. Feature weights are task-dependent, such that, for example, the features of the subject of comparison loom larger than the referent's, and common features matter more in judgments of similarity, whereas distinctive features receive greater attention in judgments of dissimilarity. This simple and elegant theory was able to explain observed asymmetries in similarity judgments (A is more similar to B than B is to A), and the fact that item A may be perceived as quite similar to item B and item B quite similar to item C, but items A and C may be perceived as highly dissimilar. Foreshadowing the immensely elegant work to come, these early papers were predicated on the technical mastery of relevant normative theories, and explored simple and compelling psychological principles until their unexpected, and often striking, theoretical implications became apparent.

Another impressive project concerned the mathematical and axiomatic foundations of measurement, in the physical sciences, but especially in the study of behaviour. Although fundamental to modern science, measurement was long considered unproblematic. In fact, it represents non-trivial issues concerning the assignment of numbers to objects in terms of their structural correspondence. Our measurement models, for example, are often are not determined by the data. Tversky's involvement in this project would stretch over two decades and result in three massive volumes (co-authored with Krantz, Luce, and Suppes 1971, 1989, 1990).

The Collaboration with Daniel Kahneman

Tversky's long and extraordinarily influential collaboration with Daniel Kahneman began in 1969 and spanned the fields of judgment and decision making. Having recognized that intuitive predictions and likelihood estimates tend not to follow the principles of statistics or the laws of probability, Tversky and Kahneman (1974) embarked on the study of biases as a method for investigating judgmental heuristics. The beauty of the work was most apparent in the interplay of psychological intuition with normative theory, accompanied by memorable demonstrations. The research showed that judgments often violate basic normative principles despite the fact that people are quite sensitive to these principles' normative appeal. An important theme in this work is a rejection of the claim that people are not able to grasp the relevant normative considerations. Rather, recurrent and systematic errors are attributed to people's reliance on intuitive judgment and heuristic processes in situations where the applicability of normative criteria is not immediately apparent. The experimental demonstrations are noteworthy not only because they violate normative theory, but also because they contradict people's own assumptions about how they make decisions.

Two early examples of judgmental heuristics illustrate this tension:

1. When presented with a description of Linda, a young, single, outspoken and very bright woman, who majored in philosophy, had participated in anti-nuclear demonstrations, and is concerned with issues of discrimination and social justice, most people think Linda is more likely to be a feminist bank teller than a bank teller – even though, of course, the likelihood of the latter must be greater than the former (since all feminist bank tellers are bank tellers).
2. When asked to estimate the number of seven-letter words on a typical page of English text, people are inclined to guess that there are fewer words whose penultimate letter is N than end in ING – even though the latter are necessarily a subset of the former.

In both cases, a heuristic judgment leads to what is known as the conjunction fallacy. In the first, people rely on the fact that Linda is more similar to a feminist bank teller than to a prototypical bank teller; in the second, frequency is judged via the ease with which examples can be brought to mind. In both cases, the reliance on intuitive heuristics leads people to ignore simple normative constraints that, upon reflection, they readily endorse.

In 1979, Kahneman and Tversky published their seminal paper on prospect theory. Although the theory is formally confined to the analysis of individual choice between binary monetary gambles, it incorporates fundamental insights that have revolutionized current theorizing about decision making more generally. Contrary to the notion of utility maximization, which focuses on final assets, the psychological carriers of value in prospect theory are gains and losses relative to some reference point, which is often the status quo. Diminishing sensitivity to greater amounts leads prospect theory's value function to be concave for gains and convex for losses (that is, above and below the reference point, respectively), yielding risk aversion for gains and risk seeking for losses (except for very low probabilities, where these trends can reverse). Because prospects can often be framed as gains or as losses relative to some reference point, this can generate 'framing effects', wherein alternative descriptions trigger opposing risk attitudes and elicit discrepant preferences regarding the same final outcomes. For example, imagine being $300 richer than you are and having a choice between $100 for sure and an equal chance at $200 or nothing. Alternatively, imagine being $500 richer and having to choose between a sure $100 loss and an equal chance to lose $200 or nothing. Although the two scenarios offer the same final outcomes ($400 versus an equal chance at $300 or $500), people tend to prefer the certain $100 gain in the first and the chance of a greater loss or nothing in the second, thus expressing opposing preferences.

According to prospect theory, people are loss averse: the loss associated with giving up a good is greater than the pleasure associated with obtaining it. Loss aversion yields 'endowment effects' wherein the mere possession of a good can lead to higher valuation of it than if it were not in one's possession (Kahneman et al. 1990), and it can create a general reluctance to negotiate or trade because the disadvantages of departing from the status quo loom larger than the advantages presented by possible alternatives (Samuelson and Zeckhauser 1988). Furthermore, the impact of probabilities in prospect theory is not linear; rather, it consists of a transformation of the relevant probabilities into 'decision weights' which capture the impact on decision makers, exhibited most clearly at the extremes of certainty and impossibility. For example, a reduction in the likelihood of a threatening outcome from .02 to 0 has a much greater impact on people (as exhibited, say, in their willingness to pay) than a comparable change in likelihoods from .67 to 65.

Later Work

Tversky returned to the study of judgment and in his work on support theory (Tversky and Koehler 1994), a theory of probabilistic judgment that formally distinguishes between events in the world and the manner in which they are mentally

represented. Probabilities in support theory are attached not to events, as in standard models, but rather to descriptions of events, called hypotheses. Probability judgments are based on the support (strength of evidence) of the focal hypothesis relative to that of alternative, or residual, hypotheses. The theory distinguishes between *explicit disjunctions,* which are hypotheses that list their individual components (for example, 'a car crash due to oil spill, or due to driver fatigue, or due to break failure'), and *implicit disjunctions* that do not ('a car crash'). According to the theory, unpacking the description of an event from an implicit to an explicit disjunction generally increases its support and, hence, the perceived likelihood. As a result, alternative descriptions of an event can give rise to substantially different judgments.

A fundamental assumption underlying normative theories is the extensionality principle: options that are extensionally equivalent are assigned the same value, and extensionally equivalent events are assigned the same probability. Normative theories are concerned with options and events in the world: different descriptions of the same states are similarly evaluated. According to Tversky's analyses, on the other hand, judgments and decisions are constructed, not merely revealed, during their elicitation, and their construction depends on the framing of the problem, the method of elicitation, and the valuations and attitudes that these trigger. The extensionality principle is deemed descriptively invalid because alternative decision contexts and alternative descriptions of options or events often produce systematically different judgments and preferences.

Behaviour, Tversky's research made clear, is the outcome of normative ideals that people endorse upon reflection, combined with psychological processes that intrude upon and shape behaviour independently of any deliberative intent. These insights led to dramatic and memorable studies concerning, among others, the hot hand in basketball (Tversky and Gilovich 1989), the perceived relationship between weather and rheumatism (Redelmeier and Tversky 1996), money illusion (Shafir et al. 1997), self-deception (Quattrone and Tversky 1984), overconfidence (Griffin and Tversky 1992), and a variety of other economic, medical and political decisions. Tversky was an intellectual giant whose work had an exceptionally broad appeal, to economists, philosophers, statisticians, physicians, political scientists, sociologists and legal theorists, among others.

Tversky taught at Hebrew University (1966–78) and at Stanford University (1978–96), where he was the inaugural Davis–Brack Professor of Behavioral Sciences and Principal Investigator at the Stanford Center on Conflict and Negotiation. He spent leave periods at Harvard University, the Center for Advanced Studies in the Behavioral Sciences, the Center for Advanced Study at Hebrew University, and the Oregon Research Institute. After 1992 he held an appointment as Senior Visiting Professor of Economics and Psychology and Permanent Fellow of the Sackler Institute of Advanced Studies at Tel Aviv University.

Tversky won many awards for diverse accomplishments. As a young officer in 1956, he earned Israel's highest honour for bravery for rescuing a soldier who had frozen in panic after lighting an explosive charge. His dissertation, under the supervision of Clyde Coombs, won the University of Michigan's Marquis Award. He won the Distinguished Scientific Contribution Award of the American Psychological Association in 1982, a MacArthur Prize in 1984, and the Warren Medal from the Society of Experimental Psychologists in 1995. He was a foreign member of the National Academy of Sciences, and a member of the Econometric Society and the American Academy of Arts and Sciences. He was awarded honorary doctorates by the University of Göteborg, the State University of New York at Buffalo, the University of Chicago, and Yale University.

Tversky was in the midst of an enormously productive time when he died of metastatic melanoma on 2 June 1996, at his home in Stanford, California. For a selection of his writings, as well as a complete bibliography, see Shafir (2004); for excellent collections of papers influenced by Tversky's work on judgment and choice, respectively, see Gilovich et al. (2001), and Kahneman and Tversky (2000).

When it awarded Daniel Kahneman the 2002 Nobel Memorial Prize in Economic Sciences 'for having integrated insights from psychological

research into economic science, especially concerning human judgment and decision-making under uncertainty', the Royal Swedish Academy of Sciences, which does not award prizes posthumously, took the unusual step of acknowledging Tversky in its Nobel citation, explaining that his joint work with Kahneman formulated alternative theories that better account for observed behaviour. Two months later, Tversky also posthumously won with Kahneman the prestigious 2003 Grawemeyer Award, which recognizes powerful ideas in the arts and sciences. The citation noted that it was 'difficult to identify a more influential idea than that of Kahneman and Tversky in the human sciences'.

See Also

- ▶ Kahneman, Daniel (Born 1934)
- ▶ Prospect Theory

Selected Works

1969. The intransitivity of preferences. *Psychological Review* 76: 31–48.

1977. Features of similarity. *Psychological Review* 84: 327–352.

1971. (With D.H. Krantz, R.D. Luce and P. Suppes.) *Foundations of measurement: Vol. 1. Additive and polynomial representations*. San Diego: Academic Press.

1974. (With D. Kahneman.) Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1131.

1979. (With D. Kahneman.) Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

1984. (With G.A. Quattrone.) Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology* 46: 237–248.

1989. (With T. Gilovich.) The cold facts about the 'hot hand' in basketball. *Chance* 2(1): 16–21.

1989. (With D.H. Krantz, R.D. Luce and P. Suppes.) *Foundations of measurement: Vol. 2. Geometrical, threshold, and probabilistic representations*. San Diego: Academic Press.

1990. (With D.H. Krantz, R.D. Luce and P. Suppes.) *Foundations of measurement: Vol. 3. Representation, axiomatization, and invariance.* San Diego: Academic Press.

1992. (With D. Griffin.) The weighing of evidence and the determinants of confidence. *Cognitive Psychology* 24: 411–435.

1994. (With D.J. Koehler.) Support theory: A nonextensional representation of subjective probability. *Psychological Review* 101: 547–567.

1996. (With D.A. Redelmeier.) On the belief that arthritis pain is related to the weather. *Proceedings of the National Academy of Sciences* 93: 2895–2896.

1997. (With E. Shafir and P. Diamond.) On money illusion. *Quarterly Journal of Economics* 112: 341–374.

2000. (With D. Kahneman, eds.) *Choices, values, and frames.* New York: Cambridge University Press/Russell Sage Foundation.

Bibliography

Gilovich, T.D., D.W. Griffin, and D. Kahneman (eds.). 2001. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Kahneman, D., J.L. Knetsch, and R. Thaler. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98: 1325–1348.

Samuelson, W., and R. Zeckhauser. 1988. Status quo bias in decision making. *Journal of Risk and Uncertainty* 1: 7–59.

Shafir, E. (ed.). 2004. *Preference, belief, and similarity: The selected writings of Amos Tversky.* Cambridge, MA: MIT Press.

Twiss, Travers (1809–1897)

Murray Milgate

In economics, Twiss's reputation rests primarily upon two contributions: one on the machinery question and the other a *View of the Progress of Political Economy in Europe since the Sixteenth*

Century (1847) of some 300 pages. Both of these works originated in lectures during his tenure as Drummond Professor of Political Economy at Oxford (1942–7). The latter numbers with McCulloch's much shorter *Historical Sketch of the Rise and Progress of the Science of Political Economy* (1926) as being among the first significant histories of the discipline published in English. The only works of comparable significance in the area which predate it appeared in French: Blanqui's *Historie de l'économie politique en Europe* (1837–8) and Jean Paul Alban de Villeneuve-Bargemon's *Historie de l'économie politique* (1836–8 and 1841). Twiss acknowledges his debt to the abovementioned authors, but has been criticized (for example, by Cossa) for a tendency to rely too heavily upon second-hand sources in the construction of his argument.

The published versions of his lectures at Oxford are all that Twiss left to the literature of economics.

Twiss was born in London on 19 March 1809 and died there on 14 January 1897, and was educated at University College, Oxford, taking his BA (in mathematics and classics) in 1830. From 1830 until 1863 he was a fellow of that college. In 1835 he commenced the study of law in Lincoln's Inn and was admitted to the Bar in 1840. Following his term as Drummond Professor (in which he succeeded Merivale), he turned more and more to the study of international law, and in 1852 he was elected to the chair in that field at King's College, London. In 1855 he moved to Oxford as Regius Professor of Civil Law, where he remained until 1870. In 1867 he became the Queen's advocate-general, and was knighted in 1868.

At this point occurred 'the catastrophe which put an end to his official career', as the original edition of this *Dictionary* put it. It seems that in 1872, Twiss instituted an action for malicious libel with intent to extort against a solicitor who had put about statements impugning the moral propriety of Twiss's wife. As the case proceeded, Lady Twiss was called to testify. However, an arduous cross-examination proved to be too much for her, and she departed London before its conclusion, thus causing Twiss's case to collapse and precipitating his resignation from all offices. Of course, it is not surprising (given the climate of the times) that Lady Twiss's breakdown should have been interpreted as telling evidence against her – but from what we now know of these extraordinary Victorian public rituals over sexual behaviour and preference, and of the pressures placed on the principal actors in such notorious trials, a rather different verdict might just as plausibly be drawn from the episode. From the point of view of individual and social psychology, however, even more interesting is the question of just why these kinds of cases were voluntarily brought before the courts in the first place.

Selected Works

1845. *On certain tests of a thriving population*. London.
1847. *View of the progress of political economy in Europe since the sixteenth century.* London: Longmans, Brown, Green and Longmans.
1861. *Law of nations considered as independent political communities.* Oxford: Oxford University Press.

Two-Part Tariffs

Charles A. E. Goodhart

Abstract

The Bank of England, founded in 1694 to finance war against France, soon became Britain's largest bank. It became responsible for maintaining the gold standard and acting as lender of last resort. To do so, it had to withdraw from commercial banking. After failing to stay on gold (1931) the Bank became subservient to the Chancellor in macro-monetary policy and was nationalized in 1946. Operational independence to set interest rates in pursuit of an inflation target was restored in the 1990s, while its previous

functions, notably bank supervision, debt management, and foreign exchange intervention, fell away.

Keywords

Bagehot, W.; Bank for International Settlements; Bank of England; Bank of Scotland; Bank rate; Banking crises; Banking supervision; Basel Committee on Banking Regulation and Supervisory Practices; Big Bang; Bretton Woods system; Bullion; Central bank independence; Central banking; Cost-push theory of inflation; Dutch East India Company; European Monetary System; Exchange controls; Exchange rate mechanism; Exchange rate targets; Financial intermediaries; Financial liberalization; Financial repression; Financial Services Authority; Friedman, M.; Gold standard; Incomes policies; Inflation; Inflation targeting; London Clearing House; Medium Term Financial Strategy (UK); Monetarism; Monetary policy; Monetary targets; Natural rate of unemployment; Phillips curve; Radcliffe Report; Royal Bank of Scotland; South Sea Company; Stagflation; Thornton, H.; Trade unions; Treasury bills; Velocity of circulation

JEL Classifications

E5

The primary motivation for the establishment of the Bank of England was the need to raise funds to help the government finance the then current war against France, although the view had also developed that a bank could help to 'stabilize' financial activity in London given periodic fluctuations in the availability of currency and credit. An original proposal by William Paterson in 1693 for a government 'fund of perpetual interest' was turned down in favour of another proposal by Paterson in 1694 to establish a company known as the Governor and Company of the Bank of England, whose capital, once raised, would be lent in its entirety to the government.

An ordinary finance act, now known as the Bank of England Act (1694), stipulated that the Bank was to be established via stock subscriptions which were to be lent to the government. A governor, deputy governor and 24 directors were to be elected by stockholders (holding d500 or more of stock).

The Evolution of the Bank's Objectives and Functions, 1694–1914

Under its original charter the Bank was allowed to issue bank notes, redeemable in silver coin, as well as to trade in bills and bullion. The notes of the Bank competed with other paper media of exchange, which comprised notes issued by the Exchequer and by private financial companies. In addition, customers could maintain deposit accounts with the Bank, which were transferable to other parties via notes drawn against deposit receipts (known as accomptable notes), thus providing an early form of cheque.

An early customer of the Bank was the Royal Bank of Scotland, which made arrangements to keep cash at the Bank from its outset in 1727. Loans were extended, predominantly in the form of discounting of bills, to individuals and companies, and the Bank undertook a large amount of lending (often via overdrafts) to the Dutch East India Company and, from 1711, to the South Sea Company. The Bank also acted as a mortgage lender, although this business never took off, and ceased some years later. Finally, an important function of the Bank was the remittance of cash to Flanders and elsewhere for the wars against Louis XIV, which was facilitated through correspondent arrangements with banks in Holland.

In 1697 the renewal of the Bank's charter for another ten years involved the passage of a second Bank Act, which increased the capital of the Bank and prohibited any other banks from being chartered in England and Wales. This monopoly was strengthened at the next renewal of the Bank's charter in 1708, when any association of six or more persons was forbidden to engage in banking activity, thereby precluding the establishment of any other joint stock banks. The Bank's position as banker to the government was consolidated in 1715 when it was decided that subscriptions for

government debt issues would be paid to the Bank, and further that the Bank was to manage the government debt (the Ways and Means Act). The Bank then acted as manager of the government's debts from that date until 1997.

The Bank also encouraged the use of its own notes in preference to other media of exchange by persuading the Treasury to increase the denomination of Exchequer bills. By 1725 the Bank's notes had become sufficiently widely used as to be pre-printed for the first time. Although a number of private banks had developed by 1750, both within and outside London, none competed seriously with the Bank in the issue of notes. By 1770 most London bankers had ceased to issue notes, using Bank of England notes (and cheques) to settle balances among themselves in what had become a well-developed clearing system. Furthermore, in 1775 Parliament raised the minimum denomination for any non-Bank of England notes to one pound and, two years later, to five pounds, effectively guaranteeing the use of Bank of England notes as the dominant form of currency. Problems relating to counterfeiting, and to the harsh treatment of those caught in the act, were, however, perennial.

In Scotland, by contrast, no note issuing monopoly existed, and banks were free to issue notes, although two banks dominated, namely, the Bank of Scotland and the Royal Bank of Scotland. Furthermore, several private note-issuing banks were in business in Ireland, and the Bank of Ireland was established in 1783. These banks relied on the Bank of England to obtain silver and gold, particularly during times of financial stress, such as 1783 and 1793.

Following a dramatic rise in government expenditures after 1793 due to the war against France, which caused a large rise in the Bank's note issue, the Bank's gold holdings fell sharply. After a scare about a French invasion convertibility was suspended in 1797, and resumed only in 1821. In view of the financial exigencies of the war, and the fact that there was in such circumstances no limit to the expansion of its note issue, now effectively legal tender, by the Bank, a privately owned company, what is in retrospect surprising about the period of suspension is how comparatively low the resulting inflation was. Even so, it was high enough to set off a major debate on its causation, for example in the Parliamentary Committee on the High Price of Bullion (1810). This period saw a further consolidation of the Bank as a note issuer, since it began to issue small denomination notes (given the shortage of silver and gold coin), which became legal tender in 1812. Furthermore, in 1816 silver coin ceased to be legal tender for small payments. The government also moved most of its accounts to the Bank in 1805 (in 1834 all government accounts were finally moved to the Bank).

During the 18th century and early part of the 19th century, smaller country banks had proliferated throughout England and Wales, many issuing their own notes. Given the prohibition on joint stock banking, the capital of these banks was usually small, and they regularly became insolvent, especially when the demand for cash (coin) became strong. This contrasted sharply with Scotland, where joint stock banking and branch banking were permitted, and relatively few failures occurred. Following a severe banking crisis in 1825, during which many English country banks failed, an Act renewing the Bank's charter (in 1826) abolished the restrictions on banking activity more than 65 miles outside of London. This led to the establishment of several joint stock banks, while the Bank countered by opening several branches throughout England.

Thus, a semblance of a banking 'system' began to emerge by 1830, with the Bank of England as the 'central' bank. By far the best book on such nascent central banking at this time was that written by Henry Thornton, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802). The practice of banks placing surplus funds with bill brokers also developed, with the Bank beginning to extend secured loans to these brokers on a more or less regular basis. In 1833 joint stock banks were finally allowed to operate in London, although they were not permitted to issue notes and thus were essentially deposit-taking banks only. The same Act specified that Bank of England notes were legal tender, and the Bank was also given the freedom to raise its discount rate freely (until then usury laws had placed a ceiling on interest rates) in

response to cash outflows. The Bank's reaction (an early reaction function), in varying its interest rate, to cash inflows and outflows became codified around this time in what became known as the Palmer rule, after Horsley Palmer, Governor 1830–33, though the rule itself is usually dated from 1827.

The position of Bank of England notes was consolidated in an important Act, passed in 1844, generally known as the Bank Charter Act, preventing all note issuers from expanding their note issue above existing levels, and prohibiting the establishment of any new note-issuing banks. The 1844 Act also separated the issue and banking functions of the Bank into different departments, and required the Bank to publish a weekly summary of accounts.

Given that it did not pay interest on its deposits, the deposit activity of the Bank could never really compete with that of other banks, which expanded rapidly from 1850 onwards. In 1854, joint stock banks in London joined the London Clearing House, and it was agreed that clearing by transfer of Bank of England notes would be abandoned in favour of cheques drawn on bank accounts held at the Bank. Ten years later the Bank of England itself entered this clearing arrangement, and cheques drawn on bankers' accounts at the Bank became considered as paid.

Although the Bank had, from the beginning of the 19th century, periodically bought or sold exchequer bills to influence the note circulation, explicit open-market borrowing operations to support its discount rate began in 1847. From 1873 until 1890 the Bank almost always acted as a borrower rather than a lender of funds, as there were typically cash surpluses. As a result, the Bank introduced the systematic issue of Treasury bills via a regular tender offer in 1877. Treasury bills had a much shorter maturity (three to twelve months) than Exchequer bills (five or more years), and were to play an important role in raising funds from the outset of the First World War onwards.

By 1890, the Bank's role as lender of last resort became undisputed when it orchestrated the rescue of Baring Brothers and Co., a bank whose solvency had become suspect, threatening to cause systemic problems. Earlier, in 1866, the failure of a discount house, Overend, Gurney and Co., had precipitated a financial panic, during which the Bank discounted large amounts of bills and extended considerable loans. The Bank, however, was criticized for not doing more to prevent the onset of such a panic, not least by Walter Bagehot in his famous book *Lombard Street* (1873).

Throughout the 19th century, the Bank streamlined its discount facilities. In 1851 it overhauled its discount rules, stipulating that only those parties having a discount account could present bills, and that these bills had to have a maturity of fewer than 95 days and be endorsed by two creditworthy firms. In the latter part of the century, however, the Bank gradually came to favour discount houses, often by presenting them with better rates of discount, and the range of firms doing discount business with the Bank declined. Discount houses were favoured because there was tension then between the Bank and the rapidly growing commercial banks – there was much banking consolidation via mergers between the 1870s and 1914 – and dealing via the intermediation of the discount houses enabled the Bank to influence market rates without having to interact directly with the joint-stock banks as counterparties.

Until the First World War the Bank pursued a discount policy which was primarily aimed at maintaining its gold reserves (as noted earlier) and which was conducted largely independently of the government. During the First World War, however, a clash occurred between the Bank Governor (Cunliffe) and the Chancellor (Law), during which the government made clear that it bore the ultimate responsibility for monetary policy, and that the Bank was expected to act on its direction.

A Subservient Bank, 1914–1992

The First World War was a major watershed not only in the history of the Bank but in the world more widely. It ushered in a half-century of increasing government intervention in every country, of a move towards socialist economies in most, and of communism in a wide swathe of

countries. Under these circumstances the Bank became increasingly subservient to the government, in practice to the Chancellor of the Exchequer and to the Treasury, in the conduct of macromonetary policy, its previous primary function.

Initially, however, there was little perception that the war and the rise of socialist ideas had irretrievably altered the context for policy. There was a desire to return to the previous regime, the gold standard, with its tried and true verities, as expressed in the Cunliffe Committee Report (the first report of the Committee on Currency and Foreign Exchange 1919). That was probably inevitable under the circumstances, but a much more questionable decision was to return at the pre-war parity (against gold) despite the war-induced loss of markets (especially for the UK's main staples, textiles, coal, and iron and steel) and of competitiveness. Several of the other belligerent states, notably France, had inflated, and allowed their exchange to float downwards by so much that they did not seek to re-peg at the previous parity, but could choose a more suitable and competitive rate. While the decision to return to gold at the pre-war parity, steadfastly supported by the Bank, has been much criticized, the modern theory of time inconsistency provides some defence, namely, if the Bank had started to change the chosen rate to suit the immediate conjuncture it would have been expected to do so again in future, making commitment to the regime less credible.

Be that as it may, conditions after the First World War, with a weak balance of payments and a massively inflated money stock and floating debt, were hardly conducive to the re-establishment of gold standard conditions. Indeed, the authorities initially felt forced to move in the other direction, to unpeg the sterling–dollar rate that had been established since 1916 and formally to leave the gold standard in March 1919. The ending of the war led then to an extremely sharp and short boom and bust, in which tight monetary policy played a major role in the subsequent deflation (see Howson 1975). From then until the return to gold at the pre-war parity of $4.86 to the pound in 1925, the Bank advocated keeping the Bank rate high enough to facilitate that regime change, but decisions on Bank rate and on the conduct of monetary policy were joint, in that no proposal by the Bank could be activated without the agreement of the Chancellor and HM Treasury; the Treasury view, however, then was in line with classical thought, namely, that monetary policy could and should impinge primarily on nominal prices, with real output affected by real factors.

Despite the boom in the USA, growth in the UK was perceived as remaining low and unemployment high, at least as compared with its main comparator countries, in the 1920s. This was in part due to the continuing problems of restoring a successful economic regime in Europe, wherein German reparations had a malign effect. Although the Bank had lost much of its power to direct domestic monetary policy (to Whitehall), the Bank and its Governor, Montagu Norman, played a leading role in the various international exercises to try to restore Europe to normality and to the gold standard, (Sayers 1976, ch. 8); and Sir Otto Niemeyer, a top Bank official, spread the gospel of establishing central banks to maintain price stability to the Dominions.

This whole structure came apart in the crisis that started in the USA in 1929 and then engulfed the rest of the world progressively through the subsequent four years. How far that collapse was itself exacerbated by the attempt to restore the gold standard has been explored by Eichengreen (1992). The UK was not in a strong economic position to avoid the world recession, but suffered a much smaller decline in output than in the USA or much of Continental Europe. The struggle to maintain the gold standard had required the maintenance of high interest rates, despite the imposition of controls on new issues in sterling by foreign governments. Despite high unemployment, wages and prices remained too sticky to allow the restoration of international competitiveness, though quite why this was so remains a debated issue.

With the gold standard collapsing in Europe and social pressures rising in the UK, there was diminishing political will to take the measures that appeared necessary to maintain the gold standard. The government decided to abandon it (in Norman's absence) in September 1931. From that moment onwards, until May 1997, the

decision to alter the Bank rate moved decisively to Whitehall, effectively into the hands of the Chancellor, advised by HM Treasury. Of course, the Bank could, and did, make suggestions and played a major role in all the discussions, but the Chancellor took the decisions. Indeed, from June 1932 until November 1951 a policy of cheap money was followed whereby Bank rate was held constant at two per cent. Norman stated in 1937, 'I am an instrument of the Treasury'.

Meanwhile, the Bank was becoming more professional. The old system of circulating the Governor's chair in turn among the directors of the Bank, who were appointed from city (but not commercial bank) institutions, was superseded by the continuing governorship of Montagu Norman from 1920 until 1944. While this arose by happenstance rather than intention (see Sayers 1976, ch. 22), it gave the Bank highly skilled, even if also highly idiosyncratic, leadership. Moreover, Norman introduced economists and other able officials into both the staff and the Court (the largely ceremonial board) of the Bank, although it is (apocryphally) recorded that Norman told one such economist, 'You are not here to tell me what to do, but to explain why I have done what I have already decided to do.'

In effect, the Bank had already become nationalized by the end of the Second World War. So the formal act of nationalization in 1946 brought about no real substantive changes, except that the Governor and his deputy (there has as yet been no woman Governor, although Rachel Lomax became the first female Deputy Governor in 2003), were appointed by the government for five years, renewable once more in most cases. Indeed, the more profound changes were brought about by Governor Gordon Richardson (1973–83) in the early 1980s. Until then, the Governor had been rather akin to a chairman, with the deputy and other internal directors as members of the board, setting strategy. Much of the executive power still lay with the Chief Cashier, who acted as leader of the heads of department, who ran the Bank. There was a clear break, a division, between the staff in the departments on the one hand and the Governors and Directors on the other. Richardson changed all that, concentrating power in the Governors' hands, sharply demoting the role of Chief Cashier, and underlining the precedence of (internal) directors over heads of department in all policy matters.

So, as power to decide the course of monetary policy – and to set the Bank rate passed to Whitehall, what did these professional central bank officials do? The Bank came to have three main areas of responsibility. The first was the management of markets, notably the money market, the bond (gilts) market and the foreign exchange market. The UK had come out of the Second World War with a massively inflated ratio of debt to GDP, and its management had remained difficult and delicate, at least until after the War Loan Conversion of 1932. No sooner, however, had debt management been thereby put on a sounder foundation than the Second World War led to a further upsurge in the debt ratio, which led once again to debt management becoming a major preoccupation of policy. Thereafter, a combination of generally prudent fiscal policies, so that the debt ratio fell steadily, and then unexpected inflation in the 1970s, which accelerated the decline in the debt ratio, and market reforms in the 1980s, enabled the procedures of debt management to become simpler and standardized. Similarly, the floating exchange rate in the 1930s, followed by attempts to maintain pegged exchange rates both during the Second World War and thereafter under the Bretton Woods system, against a background of perennially weak balance of payments conditions, made the management of the UK's foreign exchange reserves and intervention on the foreign exchange market a crucial function of the Bank until 1992, when the UK was forced out of the European exchange rate mechanism. During crises the officials in charge of such foreign exchange operations were in telephone communication with the Chancellor and, occasionally, the Prime Minister at frequent intervals.

The Bank held that such market operations required a special professional expertise (though HM Treasury remained sceptical). The Bank threw itself into such activities with enthusiasm, and defended its pre-eminent role in this respect stoutly against all outside encroachment or criticism. Indeed, its market 'savvy' was its most

powerful lever to persuade the Chancellor to its views in any debate; 'I am sorry, Chancellor, but the market will not accept that policy' was the strongest card it had to play, and that card was played often and with alacrity.

Although ultra-cheap money, with Bank rate held at two per cent, was abandoned in 1951, when the Conservative Party was returned to office, monetary policy in general, and interest rates in particular, were still seen as both more ineffective and uncertain in their impact on domestic demand than the supposedly more reliable fiscal policy, a conclusion upheld by the controversial Radcliffe Report (1959). Consequently, fiscal policy was used to try to steer domestic demand while interest rates were raised to protect the balance of payments during the regular bouts of external weakness, and otherwise held low both to ease government finance and to support fixed investment. The outcome was a system in which inflationary pressures regularly threatened both the internal and external value of the currency. The chosen solution was to supplement market measures by direct interventions, in the case of external pressure via exchange controls, in the case of monetary expansion via direct controls on bank lending to the private sector. In both instances the Bank acted as the administrative agent of HM Treasury.

Such direct controls were introduced (on bank lending), or greatly extended and tightened (exchange controls), with the onset of the Second World War in 1939, but were continued, for the reasons outlined above, until 1971 for bank lending and 1979 for exchange controls. The administration of exchange controls required a large staff, but, unlike with its market operations, the Bank had little enthusiasm for acting in this guise. The Bank hoped to restore London to its former role as an international financial centre. While it succeeded in this through its encouragement of the Eurodollar market, aided by inept US policies, the continued administration of exchange controls remained an unwelcome burden. The same was true for direct controls on bank lending. Such controls were regarded by politicians as a comparatively painless way of dampening demand and inflation, while they were resented by commercial bankers. The Bank found itself in the middle of these disputes, and grew painfully aware of such controls' stultifying effect on efficiency, dynamism and growth. The Bank, inspired by John Fforde (the then executive director in charge of domestic finance, and subsequent Bank historian), pressed hard for these controls to be dismantled, and succeeded with the liberalizing reform of Competition and Credit Control (Bank of England 1971).

As with many other cases of banking liberalization, such as in Scandinavia at the end of the 1980s, this was followed by an expansionary boom and then a bust, the fringe (secondary) bank crisis of 1973/74 (Reid 1982). While there remain questions about how monetary policy could have been better applied to prevent the prior monetary boom (1972/73), there was no question but that the financial crisis found both the Bank and the banks unskilled in risk management and unprepared for adverse shocks to financial stability. The long period of financial repression – that is, controls on bank lending to the private sector and force-feeding with government debt – had had the by-product of making the (core) commercial banking system safe between the mid-1930s and the early 1970s. The central banking function of maintaining financial stability, via regulation and supervision, had atrophied.

This had not been so earlier, and the Bank had been closely involved in the rescue of Williams Deacon's Bank by the Royal Bank of Scotland in 1930 (Sayers 1976, ch. 10), and in helping to shape the structure of both the commercial banking system and the London Discount Market Association. Williams Deacon's had got into trouble largely because of bad debts from Lancashire cotton companies. Norman, and the Bank, extended their structural interventions beyond banking to try to encourage strategic amalgamations to shore up the positions of weakened companies in a variety of industries, such as cotton, steel, shipping, armaments (Sayers 1976, ch. 14). The Bank's involvement in structural matters outside of banking itself was episodic depending on both circumstances and personalities. Another example of such Bank involvement was the considerable role it played in the reform of the UK capital market in the 1980s, more familiarly

known as 'Big-Bang'. But views on whether the Bank has any locus in such wider structural issues vary over time; the early 2000s saw a major withdrawal by the Bank from any such involvement.

The fringe bank crisis in the early 1970s was, however, a clarion call to put more emphasis on its third main function, bank supervision and regulation. The immediate result was a reorganization in the Bank. Initially a nucleus of a new specialized department was established in the Discount Office where the limited staff assigned to this role had sat, which rapidly absorbed staff and resources. Thereafter this became a separate department devoted to banking supervision and regulation (its first head was George Blunden, later to become Deputy Governor, who handed it on to Peter Cooke in 1976). Its position was regularized in the Banking Act (1979) which gave formal powers to the Bank to authorize, monitor, supervise, control and, under certain circumstances, withdraw prior authorization (tantamount to closure) for banks. No such powers had been available before that date. Meanwhile, other financial intermediaries, such as building societies or insurance companies, remained (lightly) regulated by various government departments.

The fringe bank crisis was almost entirely domestic, confined to British headquartered companies. Meanwhile, however, the onwards march of liberalization (involving the removal of direct controls, notably exchange controls in 1979) and of information technology were leading to a growing internationalization of financial business. For a variety of reasons, mostly relating to the innovation of the Eurodollar and Euro-markets, London regained its role as an international financial centre in the 1960s, and thus international monetary problems became of particular importance to the Bank, which took a leading role in such matters from the 1970s onwards.

Central bankers had met regularly at the headquarters of the Bank for International Settlements (BIS) in Basel for many years. It was, therefore, a logical step for supervisory officials also to come together at Basel on regular occasions to discuss matters of common interest. Thus was born (in 1974), as a result of an initiative from Gordon Richardson, the Basel Committee on Banking Regulation and Supervisory Practices. For the first 15 years of its existence it was chaired by the participant from the Bank of England, and was usually known by his name; thus, the Blunden Committee (1974–77) gave way in due course to the Cooke Committee (1977–88). The failures of Franklin National and Herstatt prompted the First Basel Concordat, which allocated responsibility for supervising internationally active banks to home and host authorities.

So by the mid-1970s, a need was perceived for banking supervision at both the domestic and, via consolidation, at the international levels. The purpose of these initiatives was to clarify where responsibility lay for the supervision of international banks, to prevent fragile, and possibly fraudulent, banking leading to avoidable failures and potential systemic crises.

Despite the growing number of bank supervisors, and notable success in reversing prior declines in capital ratios, the history of banking in the subsequent decades in the UK was spotted by occasional bank failures. Unlike the fringe bank crisis, none was, or was allowed to become, systemic, nor did individual depositors lose any money, except in the case of Bank of Credit and Commerce International (BCCI), and even in that case the deposit protection scheme provided some relief. The failures of Johnson–Matthey (in 1984), BCCI (in 1991) and Barings (in 1995) were all isolated cases of bad, in some respects fraudulent, banking.

The main problem of the 1970s and 1980s was, however, that of combating inflation, which soared to heights previously unknown, not only in peacetime but even in wartime, during the 1970s, up to 25 per cent per annum. There were three main theories, though divisions between them were never completely distinct. The first was the cost-push theory, that inflation was driven by overmighty trade unions, seeking to increase the relative real pay of their members; the appropriate remedy was then prices and incomes policies plus reform (and constraint) of trades unions. The second was the (vertical) Phillips curve analysis; the remedy here was to raise unemployment above the 'natural' rate to reduce inflation. The third was that inflation was a monetary phenomenon; the

remedy was to control the rate of growth of the (appropriate) monetary aggregate.

Until the mid-1970s, both major political parties, the Bank and HM Treasury all professed some combination of theories 1 (cost-push) and 2 (Phillips curve). Left-leaning politicians, academics and officials tended to put more weight on cost-push. In the 1960 and 1970s the third, monetarist, view seemed to explain events better and gained strength, not only in the USA (Milton Friedman) but also in the UK. In particular, the surge in inflation in the UK in 1973–75 followed closely behind the rapid expansion of broad (but not narrow) money in 1972–73. So, when in opposition, the leading Conservative politicians Keith Joseph and Margaret Thatcher embraced a version of monetarism.

When they came to power in 1979, they tried to commit monetary policy to follow a target for broad money, via the Medium Term Financial Strategy. In order to achieve this, nominal, and real, interest rates were kept high, and the exchange rate appreciated sharply, partly under the influence of North Sea oil and confidence in Thatcherite policies. Inflation duly declined, as planned, but broad money growth did not. This latter was partly due to the abolition of the 'corset' in 1980. The 'corset' was a reformulated, and somewhat disguised, direct control over commercial bank expansion that had been pressed into service on several occasions during the 1970s. The Bank was glad to see the end of exchange controls and direct controls over bank lending, but had never shared the government's monetarist faith in trying to set, and stick to, targets for the growth of (the various) monetary aggregates.

The empirical demonstration of the unpredictability of the relationship between (broad) money and nominal incomes in the early 1980s soon weakened the government's own faith. After moving from one monetary target to several joint targets, and an attempt to hit the broad money target by 'overfunding', an exercise criticized by many as artificial, the government abandoned its monetary targetry in 1986.

That left the question of how monetary policy, and with it control of inflation, was to be managed or, in the standard phrase, 'anchored'. The then Chancellor, Nigel Lawson, wanted to 'anchor' by joining the exchange rate mechanism (ERM) of the European Monetary System and leaving the steering of monetary policy to the Bundesbank. The Prime Minister, Mrs Thatcher, and her adviser, Alan Walters, were opposed, both on economic grounds (that such a pegged system was 'half-baked') and for wider political reasons. There was a battle royal in which the Bank was left on the sidelines. Lawson was sacked, but eventually Mrs Thatcher was, grudgingly, persuaded to allow the UK to join the ERM in October 1990.

This was in the aftermath of German reunification, and the expenditures connected with that led the Bundesbank to keep interest rates higher than was tolerable for the UK (or Italy). The UK was in the throes of a sharp downturn in housing prices, following an unstable housing boom in the late 1980s. With the Conservatives having become politically weaker, there was just no stomach to raise interest rates to the levels necessary to sustain the ERM. The UK was forced out in September 1992.

Independent and Focused, 1992–

The ejection of the UK from the ERM left the government and HM Treasury with the recurrent problem of how to manage, to 'anchor', monetary policy. Both monetary and exchange rate targets had been tried, and both had been found wanting. While the economic experience of the 1980s was better than that of the stagflationary 1970s, it was hardly stellar, with a boom–bust cycle at the end of the decade.

Meanwhile, a new approach had been adopted in New Zealand, whereby the central bank was given administrative freedom to vary interest rates for the purpose of hitting a target for the inflation rate, jointly set by the government and the central bank: that is, inflation targetry. This obviated one of the shortcomings of monetary targetry, namely, the unpredictability of the velocity of money; it left setting the goals of policy, the overall strategy, in the hands of government, but shifted the (constrained) discretion to vary interest rates to

the professional and technical judgement of the central bank. This procedure soon generated a strong body of academic support (for example, Fischer 1994).

Although Conservative Chancellors (both Lawson and Lamont) had toyed with the idea of giving the Bank operational independence, consecutive Prime Ministers (Thatcher and Major) refused, primarily on political grounds. Nevertheless Lamont wanted to move to an inflation target. But there was a problem of governmental credibility. To foster credibility, Lamont now encouraged (in 1992/93) the Bank to prepare and to publish an independent forecast of the likely projection for inflation, the *Inflation Report* (on the assumption of unchanged policies); this was a reversal of prior habits whereby HM Treasury and Ministers customarily censored Bank publications and discouraged any publication of internal Bank forecasts. The process of gradually giving the Bank a more independent role in setting monetary policy took a step further when the next Chancellor, Clarke, not only held a meeting with the Governor, and the Bank, to discuss future changes in interest rates, but published the minutes of the meeting, including the Governor's initial statement, verbatim; this was termed the Ken (Clarke) and Eddie (George) show. That said, Clarke had strong views on the appropriate policy and on a couple of occasions overruled the Governor's suggestions.

At that time – the mid-1990s – there were still question marks over the Labour Party's ability to manage the economy; financial markets are inherently suspicious of left-leaning governments. So Labour had more to gain (than the Conservatives), in terms of confidence and lower interest rates, by granting operational independence (back) to the Bank. In advance of the 1997 election the then shadow Chancellor, Gordon Brown, was cautious; while indicating general support for both inflation targetry and operational independence, he stated that he wanted time to see how well the Bank performed before granting such independence. But, within days of winning the election, he made that strategic change to the monetary regime.

This was, of course, a great prize for the Bank, but it did not come without cost. In the same month as operational independence was awarded to the Bank, both debt management and banking supervision were hived off, to a separate Debt Management Office (DMO) and Financial Services Authority (FSA) respectively. With the government debt to GDP ratio having declined and capital markets strengthened, debt management had become more of a routine and standardized exercise. Nevertheless, its departure to the DMO, and the fact that the float of the exchange rate after 1992 was kept 'clean', that is, without intervention, meant that much of the market operations which had been so central to the Bank in the post-Second World War period disappeared, though its money market operations, of course, continued. The administration of direct controls had gone at the beginning of the 1980s. And now banking supervision was also taken away. This meant that almost *all* the prime functions that the Bank had undertaken in its post-Second World War period of subservience had now gone. Instead, the Bank was now focused on varying interest rates to achieve the inflation target set for it by the Chancellor.

There are numerous arguments, quite evenly balanced, for whether bank supervision should be kept within a central bank or put with a separate Financial Services Authority (FSA), covering both banks and other financial intermediaries (see Goodhart 2000). Be that as it may, there are various aspects of the financial system, such as oversight of the payments' system, and of crisis management, such as lender of last resort functions, which cannot be delegated to an FSA. Moreover, the achievement of price stability is likely to be seriously compromised by any serious bout of financial instability – and vice versa, with financial stability adversely affected by price instability. So the removal of individual bank supervision does not absolve the Bank from concern with financial stability issues more widely; indeed, the Bank is specifically charged with maintaining overall systemic stability in the financial system. But exactly what that means when responsibility for the conduct of individual bank supervision is located elsewhere is not yet entirely clear.

What it certainly does mean is that the FSA, the Bank, and the political authorities as the ultimate source of any needed fiscal support have to work

extremely closely together, in advising on any new regulations (whether domestic or international), in monitoring developments (as in the Financial Stability Review), and in crisis management. This latter task would be done via the Tripartite Standing Committee (FSA, Bank, and HM Treasury), set up in 1997, although so far no such financial (as contrasted with simulated 'war games') crisis has occurred, though the Committee did meet after the terrorist attacks on 7 July, 2005. How successful crisis management by such a committee may be has yet to be seen.

The monetary policy function of the Bank, now its central preoccupation, has, however, been very successful by all the usual criteria. In several papers Luca Benati (for example, Benati 2005) has demonstrated that the variance of both GDP and of inflation around its target has been lower under the inflation targetry regime (whether taken as starting in 1992 or in 1997) than under any previous historical regime. The procedures of having a Monetary Policy Committee consisting of five senior Bank officials and four outside experts (appointed by the Chancellor), with the Committee serviced by Bank staff, has worked generally smoothly and well. So the Bank's reputation and credibility have rarely been higher, although now tightly focused on one main function.

See Also

- ▶ Banking Crises
- ▶ Bullionist Controversies (Empirical Evidence)
- ▶ Gold Standard
- ▶ Inflation Targeting
- ▶ Monetary Policy, History of

Bibliography

Acres, W. 1931. *The Bank of England from within*. London: Oxford University Press.

Andréadès, A. 1909. *A history of the Bank of England*. London: P. S. King and Sons.

Bagehot, W. 1873. *Lombard street*. London: Kegan, Paul and Co.

Bank for International Settlements. 1963. Bank of England. In *Eight European central banks*. Basle: Bank for International Settlements.

Bank of England. 1971. *Competition and credit control*. London: Bank of England.

Benati, L. 2005. The inflation-targeting framework from an historical perspective. *Bank of England Quarterly Bulletin* 45(2): 160–168.

Bowman, W. 1937. *The story of the Bank of England: From its foundation in 1694 until the present day*. London: Herbert Jenkins.

Chapham, R. 1968. *Decision making: A case study of the decision to raise the bank rate in September 1957*. London: Routledge and Kegan Paul.

Clapham, J. 1944. *The Bank of England: A history*. Cambridge: Cambridge University Press.

Clay, H. 1957. *Lord Norman*. London: Macmillan.

Committee on Currency and Foreign Exchange After the War (Cunliffe Committee). 1918. *First Interim Report*, Cmnd. 9182; and 1919. *Final Report*, Cmnd 464. London: HMSO.

Eichengreen, B. 1992. *Golden fetters: The gold standard and the great depression*. New York: Oxford University Press.

Feavearyear, A. 1963. *The pound sterling: A history of English money*, 2nd edn, rev. E. Morgan. Oxford: Clarendon.

Fforde, J. 1992. *The Bank of England and public policy 1941–1958*. Cambridge: Cambridge University Press.

Fischer, S. 1994. Modern central banking. In *The future of central banking*, ed. F. Capie, C. Goodhart, S. Fischer and N. Schnadt. Cambridge: Cambridge University Press.

Geddes, P. 1987. *Inside the Bank of England*. London: Boxtree.

Giuseppi, J. 1966. *The Bank of England: A history from its foundation in 1694*. London: Evans Brothers Limited.

Goodhart, C. 2000. *The organisational structure of banking supervision,* Special paper, no. 127. London: Financial Markets Group Research Centre, London School of Economics. Subsequently published in *Economic Notes* 31: 1–32.

Hennessey, E. 1992. *A domestic history of the Bank of England 1930–1960*. Cambridge: Cambridge University Press.

Howson, S. 1975. *Domestic monetary management in Britain, 1919–38*. Cambridge: Cambridge University Press.

Radcliffe Report. 1959. *Report: Committee on the working of the monetary system*, Cmnd 827. London: HMSO.

Reid, M. 1982. *The secondary banking crisis, 1973–75: Its causes and course*. London: Macmillan.

Richards, R. 1929. *The early history of banking in England*. London: Frank Cass and Co.

Rogers, J. 1887. *The first nine years of the Bank of England*. Oxford: Clarendon.

Sayers, R. 1936. *Bank of England operations, 1890–1914*. London: P.S. King and Son.

Sayers, R. 1957. *Central banking after Bagehot*. Oxford: Clarendon.

Sayers, R. 1976. *The Bank of England, 1891–1944*. Cambridge: Cambridge University Press.

Smith, V. 1936. *The rationale of central banking*. London: P.S. King and Son.
Steele, H., and F. Yerbury. 1930. *The old bank of England*. London: Ernest Benn.
Stockdale, E. 1967. *The Bank of England in 1934*. London: Eastern Press.
Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain*. New York: Kelley, 1962.
Ziegler, D. 1990. *Central Bank, peripheral industry: The Bank of England in the provinces 1826–1913*. London: Leicester University Press.

Two-Sector Models

H. Uzawa

Keywords
Marginal rate of substitution; Rybczynski theorem; Two-sector models

JEL Classifications
D01

Problems of interest in economic theory, from both the theoretical and policy points of view, occur only when there exists a multitude of goods and services, each of which is either produced by different technologies or utilized for different purposes in consumption. However, it is often the case that an economic system with a multitude of goods and services is too complicated to analyse effectively and to derive conclusions of any practical use. two-sector models enable us to bring forth essential elements of the economic mechanisms in a more complicated real world while still making it possible to analyse graphically the basic structure of equilibrium and to understand the policy implications within the framework of the two-sector analysis. The two-sector analysis plays a particularly important role in trade theory and in growth theory.

A typical two-sector model concerns itself with an economy in which there exist two productive sectors, to be referred to as sector 1 and sector 2, respectively. In the context of growth theory, one sector produces consumption goods and the other investment goods. Both goods are assumed to be composed of homogeneous quantities and to be produced by two factors of production, capital and labour. Both capital and labour are also assumed to be composed of homogeneous quantities.

In each sector, production is assumed to be subject to constant returns to scale and diminishing marginal rates of substitution between capital and labour. Joint products are excluded and external (dis-)economies do not exist. The output in each sector is determined by the quantities of capital and labour allocated to that sector. In sector j, let Y_j be the quantity of good j produced by the input of capital and labour by the quantities K_j and L_j, respectively, then we may write

$$Y_j = F_j(K_j, L_j), j = 1, 2. \tag{1}$$

For each j, the production function $F_j(K_j, L_j)$ is linear homogeneous and continuously differentiable, so that the marginal rate of substitution between capital and labour is well defined.

Let K and L be the quantities of capital and labour which exist in the economy at a particular moment of time. If both capital and labour are assumed to be freely transferred from one sector to another and both are fully employed, then we have

$$K_1 + K_2 = K, \quad L_1 + L_2 = L. \tag{2}$$

In a typical two-sector model, it is often assumed that the allocation of two factors of production is perfectly competitive, so that in each sector the wage w is equal to the marginal product of labour and the rentals r of capital goods to the marginal product of capital:

$$w = pj\frac{\partial F_j}{\partial L_j}, \quad r = pj\frac{\partial F_j}{\partial F_j}, \tag{3}$$

where pj is the price of good j.

In what follows, good 1 is taken as the numéraire, so $p_1 = 1$ and $p_2 = p$.

Since production is assumed to be subject to constant returns to scale, the model is reduced to one involving per capita quantities only. Let us introduce the following notation:

$k = K/\mathrm{L}$: the capital–labour ratio in the economy as a whole,
$k_j = K_j/L_j$: the capital–labour ratio in sector j,
$Y_j = Y_j/L$: output of good j per capita,
$v_j = L_j/\mathrm{L}$: the proportion of labour allocation in sector j,
$\omega = w/r$: the wage–rental ratio.

The relations (1)–(3) are then reduced to the following:

$$y_j = f_j(\mathrm{k}_j)v_j, \tag{4}$$

where $f_j(k_j) = F_j(k_j, 1)$,

$$v_1 = \frac{k - k_2}{k - k_2}, \quad v_2 = \frac{k_1 - k}{k_1 - k_2}, \tag{5}$$

$$w = \frac{f_j(k_j)}{f_j'(k_j)} - k_j, \tag{6}$$

$$P = \frac{f_1'(k_1)}{f_2'(k_2)}, \tag{7}$$

$$y_1 = f_1(k_1)\frac{k - k_2}{k - k_2}, \quad y_2 = f_2(k_2)\frac{k_1 - k}{k_1 - k_2}. \tag{8}$$

The relation (6) means that the wage–rentals ratio ω is equal to the marginal rate of substitution between capital and labour. The capital–labour ratio k_j which satisfies (6) is uniquely determined for given wage–rentals ratio ω; it may be written $k_j = k_j(\omega)$, which is referred to as the optimum capital–labour ratio corresponding to the wage–rentals ratio ω. It is easily seen that the optimum capital–labour ratio $k_j(\omega)$ is an increasing function of the wage–rentals ratio ω. In fact, by differentiating (6) with respect to ω, we get

$$\frac{\mathrm{d}k_j}{\mathrm{d}\omega} = \frac{\left[f_j'(k_j)\right]^2}{f_j(k_j)f_j''(k_j)} > 0, \tag{9}$$

because of the diminishing marginal rate of substitution condition: $f_j'(k_j) > 0$ and $f_j''(k_j) < 0$.

The relationships between the price ratio p and the wage–rentals ratio ω may be obtained by differentiating (7) logarithmically, and noting (9):

$$\frac{1}{p}\frac{\mathrm{d}p}{\mathrm{d}\omega} = \frac{1}{\omega + k_2(\omega)} - \frac{1}{\omega + k_1(\omega)}. \tag{10}$$

Hence, we have the following proposition:

The relative price p of good 2 is an increasing or decreasing function of wage–rental ratio ω according to whether good 1 is more or less capital-intensive than good 2.

In particular, if good 1 is always more capital-intensive than good 2, then the relative price p of good 2 (with respect to the price of good 1) is an increasing function of the wage–rentals ratio ω. The latter is indeed nothing but the essence of the factor–price–equalization theorem. In this case, we can see from (8) that, as the endowment ratio k is increased, the output of good 1 is increased and that of good 2 is decreased, provided that the wage–rentals ratio ω or relative price p remains constant.

The wage–rentals ratio ω or price ratio p is determined once the demand conditions are specified.

The allocation of capital and labour between two sectors in the perfectly competitive situation, as described above, may be viewed from another point of view. It is easily seen that the allocation of capital and labour which satisfies (1)–(3) is nothing but the solution of the following optimum problem:

Find the allocation (k_1, k_2, L_1, L_2) which maximizes the national product

$$Y = P_1 Y_1 + P_2 Y_2$$

subject to the constraints (1) and

$$k_1 + k_2 \le K, \quad L_1 + L_2 \le L. \tag{2'}$$

In fact, wage w and rentals r are the Lagrange multipliers associated with the constraints (2′).

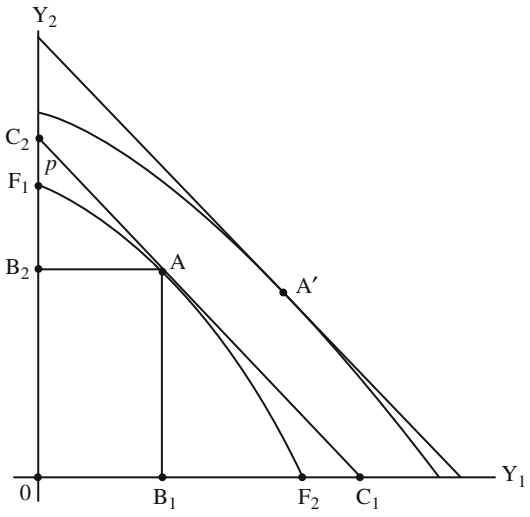
The set of all combinations (Y_1,Y_2) of two goods satisfying (1) and (2′) then is a convex set, of all possible combinations of the quantities of two goods which can be produced from the given endowments of capital and labour, K and L. The competitive allocations of capital and labour then result in those combinations of two goods for

which the national product, evaluated at prices P_1 and P_2, is maximized.

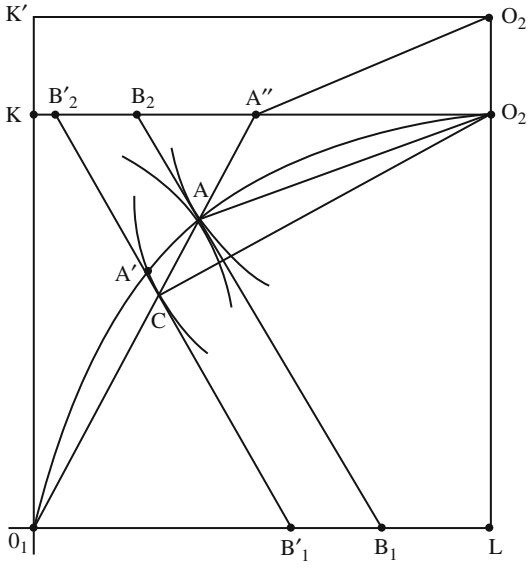
These observations lead us to the following conclusion. Namely, if the demand conditions are those obtained by an optimization of a certain community preference ordering, then the equilibrium prices and outputs are uniquely determined.

The analysis may be carried out in terms of a geometric presentation. For the given techniques of production and the given factors of production, the set of all possible combinations of two goods produced in the two-sector economy is represented by the production possibility set, as shown by the shaded area in Fig. 1. In Fig. 1, the quantities of good 1 and good 2 are measured along the abscissa and ordinate, respectively, and the boundary curve of the production possibility set is the transformation curve F_1F_2, showing the maximum quantity of one good that can be produced, given a specific quantity of the other to be produced. The transformation curve is concave toward the origin and the tangent at each point on the transformation curve has a slope equal to the relative price of two goods, as shown in Fig. 1. It is possible to prove these properties by using the contract box, as in Fig. 2. In Fig. 2, the endowments of capital and labour are measured along the sides of the box, and the allocations of capital and labour between two sectors are entered in the box from opposite corners. An efficient allocation of factors of production is realized only at a point at which two isoquants are tangent to each other. The efficient locus in the contract box corresponds to the transformation curve in Fig. 1. The configuration described in Fig. 2 represents the case where good 1 is more capital-intensive than good 2. Let the point A in Fig. 1 correspond to the point A in the contract box in Fig. 2. Suppose the quantity of good 1 to be produced is reduced by ΔY_1 and the production of good 2 is increased by ΔY_2, resulting in a shift from A to A' along the efficiency locus. At point A, the isoquants in two sectors have a common tangent; let B_1 and B_2 be the points on the labour-side at which the tangent line at A intersects. The two distances, O_1B_1 and O_2B_2, measure the values of the two goods produced, p_1Y_1/w and p_2Y_2/w, respectively. Let C be the point at which O_1A intersects with the isoquant passing through A', and let B'_1 and B'_2 be the points on the labour-side at which the tangent line at C intersects. Then $B'_1B_1 = p_1\Delta Y_1/w$ and $B'_2B_2 \simeq p_2\Delta Y_2 = w$, where the symbol $\simeq$ indicates that both sides are approximately equal, converging to the equality as ΔY_1 approaches to 0. Hence, $\Delta Y_1/\Delta Y_2 \simeq p_2/p_1$.

As the output of good 1 is reduced, the point A moves towards O_1 along the efficiency locus.



Two-Sector Models, Fig. 1 The transformation curve

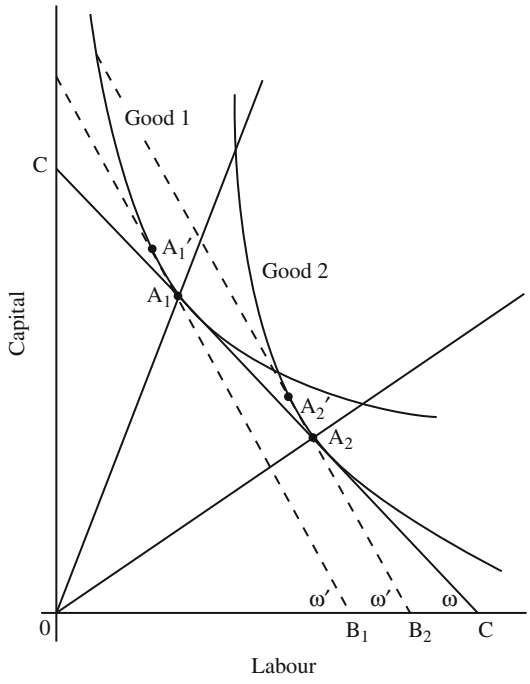


Two-Sector Models, Fig. 2 The contract box

T

Then the optimum capital–labour ratio, which is represented by LAO_1B_1, is increased, and the tangent line at A' is steeper than that at A, indicating an increase in the price ratio p_2/p_1. The transformation curve thus is shown to be concave towards the origin and the tangent at each point on it is equal to the price ratio.

The relationships between the price ratio and the wage-rentals ratio may also be discussed in terms of a two-dimensional diagram, as in Fig. 3. Suppose good 1 is more capital-intensive than good 2. For the given wage–rentals ratio ω, the unit by which good 2 is measured is so adjusted that $p_2/p_1 = 1$ and the unit-isoquants for both goods share the same cost line CC, as shown in Fig. 3. The distance OC along the abscissa measure $p_1/w = p_2/w$. Suppose the wage–rentals ratio is increased from ω to ω'. Then at the new wage–rentals ratio ω', the configuration of the cost lines in two sectors must be of the form described in Fig. 3. Namely, $OB_1 < OB_2$; hence, $p_1'/w' < p_2'/w'$, implying $p_2'/p_1' < 1 = p_2/p_1$. Thus we have proved that, as the wage–rentals ratio is increased, the price of a good which is more labour-intensive has been increased relative to that of a less labour-intensive good.



Two-Sector Models, Fig. 3 The relationships between price ratio p and wage–rentals ratio ω

The effect of an increase in the endowment of either capital or labour may also be analysed in terms of the contract box. Suppose the endowment of capital is increased from K to K' so that the new contract box is indicated by $O_1LO'_2K'$ in Fig. 2. If the relative price $p = p_2/p_1$ remains unchanged, then the factor price ratio $\omega = w/r$ also remains unchanged. Let A'' be the point at which the extension of O_1A intersects with the line originating from O'_2 which is parallel to O_2A. Then at the new configuration, the output of good 1 is increased by AA'', while the output of good 2 is decreased from O_2A *to* O'_2A'. Thus we have shown the proposition, known as the Rybczynski theorem: An increase in the endowment of capital increases the output of a good which is more capital-intensive than the other, while decreasing the output of another good which is less capital-intensive, provided the price ratio of the two goods remains constant. The resulting shift in the transformation curve is described in Fig. 3, where the efficient point A moves to A' in the new environment.

See Also

► Factor Price Frontier

Two-Sided Markets

M. Armstrong and J. Wright

Abstract

A growing number of industries are organized as so-called two-sided markets in which platforms enable interactions between two groups of users, each of which cares about the size and attributes of the other group on the same platform. The literature to date examines how platforms set prices to the two sides and whether

the resulting price structure results in market failure. The answers to these questions depend on the nature of cross-group and own-group externalities, the types of fees possible (membership or per-transaction), and whether one or both sides multihome.

Keywords

Access pricing; Advertising; Competitive bottlenecks; Complementary goods; Cross-group externalities; Matching; Multihoming; Network effects; Platforms; Price discrimination; Shopping malls; Singlehoming; Two-sided markets; User fees

JEL Classification

L10; D40; L11; L13; L15

There are many examples of markets where two or more groups of participants interact via 'platforms'. Of course, there are countless examples where firms compete to supply two or more groups. However, in a set of interesting cases, cross-group network effects are important, and the benefit enjoyed by a member of one group on a platform depends upon how well that platform does in attracting custom from the other group. For instance, a general purpose credit card scheme cannot offer a valuable service to either side unless it persuades a large number of consumers to carry its card and a large number of retailers to accept its card. The literature on two-sided markets investigates such markets.

Many examples of two-sided markets involve platforms that mediate transactions between consumers and retailers (or sellers). Examples include shopping malls, supermarkets, and debit and credit card payment schemes. Another set of examples includes matching agencies, such as real estate agencies that facilitate search and trade between home buyers and home sellers (or landlords and tenants). Advertising media including Yellow Pages, newspapers, television, and Internet portals also help match potential buyers with sellers, although in a less directed way. Software platforms such as video games, computer operating systems and word processors that connect users and application developers (or readers and writers) provide a further set of examples.

Early theoretical works on two-sided markets tended to focus on specific industries, such as Baxter's (1983) normative analysis of the structure of fees in a credit card network (see credit card industry). More general theoretical frameworks to analyse two-sided markets have been offered by Armstrong (2006), Armstrong and Wright (2007), Caillaud and Jullien (2001, 2003), Hagiu (2005), and Rochet and Tirole (2003), among others. These models extend the earlier literature on network externalities to incorporate heterogeneous agents (the two sides of the market), as well as allowing for price discrimination across the two types of users. A central question is: What determines the structure of prices in two-sided markets? For instance, why is it that shopping malls offer free parking to shoppers and recover the cost from retailers, and why does American Express charge merchants but provide rebates to cardholders? A second question concerns whether the resulting price structure causes any form of market failure. We consider both questions.

Effects of Cross-Group Externalities

Consider a generic two-sided market with buyers and sellers. Positive cross-group externalities have two effects. First, like network effects in one-sided markets, they make demand more sensitive to price. Second, like pricing for complementary goods, they make platforms charge less to one group if this increases the demand from the other group and the other group generates a positive margin. This implies that platforms will charge buyers less and sellers more when sellers value buyers more than vice versa (Armstrong 2006). By attracting buyers (with a discount), platforms can then attract the more lucrative sellers. Consistent with this idea, Yellow Pages directories are typically given away to readers for free, and profits are made entirely from charging advertisers. This assumes, of course, that the two sides cannot easily internalize the externalities

between themselves – a precondition for the structure of prices to be non-neutral in a two-sided market (Rochet and Tirole 2006). It also assumes that agents make decisions about which of the competing platforms to join. This latter assumption is explored in the following two sections.

Membership Fees Versus Usage Charges

Platforms may charge for their services on a 'lump-sum' basis: magazines set cover prices and advertisement rates, nightclubs set entry fees for men and women. Alternatively, charges might be levied on a 'per-transaction' basis: typically credit card holders receive a percentage rebate on the amount they spend, and retailers pay a percentage of the revenue they collect; real estate agents' fees are levied only in the event of a sale; charges for telecommunications service are levied on a per-call basis. And sometimes a combination of the two approaches is used: video game platforms charge consumers a fixed charge for their consoles, but game developers pay royalties for their sales.

The analysis of two-sided markets can be quite sensitive to the nature of pricing. With lump-sum charges, profits are the sum of profits obtained from each side, and it is possible to think of one side as subsidizing the other. When cross-group network effects are strong, there are often multiple consistent demand outcomes for a given set of prices, which means that some method (such as specifying consumers' beliefs about other consumers' choices) is needed to pin down demand as a function of prices uniquely (Caillaud and Jullien 2001, 2003).

With per-transaction charges, by contrast, an agent's decision whether to join a platform is less sensitive to his beliefs about the number of agents from the other side who join the platform. In this case, the equilibrium structure of prices will primarily reflect the need to balance the two sides of the market (Schmalensee 2002; Wright 2004). A matching market with lots of potential sellers but few potential buyers will not generate many successful transactions, and this suggests that more buyers need to be attracted to the platform by charging sellers rather than buyers for successful transactions.

A crucial difference between the two forms of tariff is that cross-group externalities are less important with usage charges. Charging on a usage basis is a good strategy for an entrant. If an agent has to pay a new platform only in the event of a successful transaction, then the agent does not have to worry about how well the entrant does in its dealings with the other side of the market. With per-transaction charging, to attract one side a new platform does not have to first get the other side 'on board'.

Singlehoming Versus Multihoming

When an agent chooses to use only one platform, that agent is said to 'singlehome' and when he uses several platforms he 'multihomes'. For instance, while shoppers may tend to only shop at their nearby shopping mall (singlehome), retailers may locate in several shopping malls (multihome) in order to gain access to the full range of local shoppers. Similarly, people might read a single newspaper each day, while advertisers have to place adverts in several newspapers to reach the whole readership. This pattern, where 'buyers' singlehome and 'sellers' multihome, characterizes a number of two-sided markets. Armstrong and Wright (2007) show this pattern arises endogenously when only buyers view the platforms as differentiated (as may be true in the two examples just mentioned). This leads to a 'competitive bottleneck' – platforms compete aggressively to sign up buyers, charging them less than cost (perhaps nothing), and then make their profits from sellers who want to reach these buyers and do not have a choice of which platform to join in order to reach them. Platforms need to compete to attract the singlehoming buyers but they hold a monopoly position when they deal with sellers. As charges to sellers will be too high, there will be too few sellers from a social welfare point of view. The same logic is also seen for mobile telephony (Armstrong 2002; Wright 2002), in which fixed-line callers are charged high fees to call mobile subscribers, who join a

single mobile network and receive handset subsidies for doing so. As a result, there may well be too few calls made to each mobile subscriber. To counter this market failure, fixed-to-mobile termination charges are regulated in several countries.

Negative Own-Group Externalities

In many examples of two-sided markets, agents not only like more agents from the other side, but they dislike more agents from the same side (for example, firms generally dislike advertising by rival firms in the same directory). Negative own-group externalities will justify higher prices, reflecting the 'pollution effect' of attracting additional agents. One case where negative own-group externalities arise endogenously is when sellers join the platform to better compete for buyers on the other side. Consider, for example, the case of payment schemes that attract cardholders and merchants. To the extent merchants accept cards to attract customers from each other, their private willingness to accept cards includes the surplus their customers get from using cards. As a result, card schemes will charge merchants more and cardholders less (Wright 2004). In addition, since the surplus of cardholders is over-represented in the profits of the card schemes, merchants will tend to be charged too much and cardholders too little.

Negative own-group externalities can also explain agreements to exclude rival agents from the same platform. For instance, a shopping mall may restrict the number of competing retailers (such as bookstores). If the platform finds it difficult (or costly) to recover revenue from the consumer side, this may be a way to drive up revenue from retailers. If a television channel cannot charge viewers, it may maximize profits by promising one car maker that it will not show an advert from a rival car maker in the same advertising slot. More generally, platforms act somewhat like regulators (Hagiu 2005; Rochet and Tirole 2006). They impose rules, conditions and prices for the platform that help solve various inefficiencies, at the cost perhaps of introducing others.

See Also

▶ Credit Card Industry

Bibliography

Armstrong, M. 2002. The theory of access pricing and interconnection. In *Handbook of telecommunications economics*, ed. M. Cave, S. Majumdar, and I. Vogelsang. Amsterdam: North-Holland.

Armstrong, M. 2006. Competition in two-sided markets. *RAND Journal of Economics* 37: 668–691.

Armstrong, M., and J. Wright. 2007. Two-sided markets, competitive bottlenecks and exclusive contracts. *Economic Theory* 32: 353–380.

Baxter, W.F. 1983. Bank interchange of transactional paper: Legal perspectives. *Journal of Law and Economics* 26: 541–588.

Caillaud, B., and B. Jullien. 2001. Competing cybermediaries. *European Economic Review* 45: 797–808.

Caillaud, B., and B. Jullien. 2003. Chicken & egg: Competition among intermediation service providers. *RAND Journal of Economics* 34: 309–328.

Hagiu, A. 2005. Two-sided platforms: Pricing and social efficiency. Harvard Business School mimeo.

Rochet, J.-C., and J. Tirole. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1: 990–1029.

Rochet, J.-C., and J. Tirole. 2006. Two-sided markets: A progress report. *RAND Journal of Economics* 37: 645–667.

Schmalensee, R. 2002. Payment systems and interchange fees. *Journal of Industrial Economics* 50: 103–122.

Wright, J. 2002. Access pricing under competition: An application to cellular networks. *Journal of Industrial Economics* 50: 289–315.

Wright, J. 2004. Determinants of optimal interchange fees in payment systems. *Journal of Industrial Economics* 52: 1–26.

T

Two-Stage Least Squares and The *K*-Class Estimator

N. E. Savin

Abstract

Two-stage least squares has been a widely used method of estimating the parameters of a single structural equation in a system of linear

simultaneous equations. This article first considers the estimation of a full system of equations. This provides a context for understanding the place of two-stage least squares in simultaneous-equation estimation. The article concludes with some comments on the lasting contribution of the two-stage least squares approach and more generally the future of the identification and estimation of simultaneous-equations models.

Keywords

Asymptotic distribution; Bayesian method-of-moments approach; Cowles Foundation; Full and limited information methods; Full-information maximum likelihood; Generalized least squares; Generalized method of moments; Heteroskedasticity and autocorrelation; Homoskedasticity; Identification; Indirect least squares; Instrumental variables; *k*-class estimators; Limited information maximum likelihood; Linear models; Maximum likelihood; Ordinary least squares; Reduced-form equations; Simultaneous equations models; Structural parameters; Two-stage least squares (2SLS); Two-stage least squares estimator and the *k*-class estimator; Vector autoregressions

JEL Classifications

C1

Two-stage least squares (2SLS) was originally proposed as a method of estimating the parameters of a single structural equation in a system of linear simultaneous equations. It was introduced more or less independently by Theil (1953a, 1953b, 1961), Basmann (1957) and Sargan (1958). The early work on simultaneous equations estimation was carried out by a group of econometricians at the Cowles Foundation. This work was based on the method of maximum likelihood. In particular, Anderson and Rubin (1949, 1950) developed the limited information maximum likelihood (LIML) estimator for the parameters of a single structural equation. Anderson (2005) gives the history of 2SLS a revisionist twist by pointing out that Anderson and Rubin (1950) indirectly includes the 2SLS estimator and its asymptotic distribution. The notation of that paper is difficult and the exposition is somewhat obscure, which may explain why few econometricians are aware of its contents. See Farebrother (1999) for additional insights into the precursors of 2SLS.

2SLS was by far the most widely used method in the 1960s and the early 1970s. The explanation involves both the state of statistical knowledge among applied econometricians and the state of computer technology. The classic treatment of maximum likelihood methods of estimation is presented in two Cowles Commission monographs: Koopmans (1950), *Statistical Inference in Dynamic Economic Models*, and Hood and Koopmans (1953), *Studies in Econometric Method*, which was directed at a wider audience. Among applied econometricians, relatively few had the statistical training to master the papers in these monographs, especially Koopmans (1950). By the end of the 1950s computer programs for ordinary least squares were available. These programs were simpler to use and much less costly to run than the programs for calculating LIML estimates. Owing to advances in computer technology, and, perhaps, also the statistical background of applied econometricians, the popularity of 2SLS started to wane towards the end of the 1970s. In particular, the difficulty of calculating LIML estimates was no longer an important constraint.

This article first considers the estimation of a full system of equations and then focuses on 2SLS. This approach provides a context for understanding the place of 2SLS in simultaneous-equation estimation. The article is organized as follows. A two-equation structural form model with normal errors and no lagged dependent variables is introduced in section "The Model". Section "Ordinary Least Squares" reviews the properties of the ordinary least squares estimator of the parameters of a structural equation. The indirect least squares estimator is introduced in section "Indirect Least Squares". In section "Indirect Feasible Generalized" presents the indirect feasible generalized least squares estimator, and briefly discusses maximum likelihood methods. Section "Two-Stage Least Squares"

develops two rationales for the 2SLS procedure, and the *k*-class family of estimators is defined in section "The K-Class Family". Finite sample results on the comparisons of estimators are reported in section "Finite Sample Distributions", and the concluding comments are in section "Epilogue". (Our exposition of structural-form estimation draws heavily on the treatment by Goldberger 1991. For the presentation of GMM and more recent methods of simulation-equation estimation, see Mittelhammer et al. 2000.)

The Model

In the spirit of Goldberger (1991), we consider a two-equation demand and supply model to fix ideas and notation. The endogenous variables are y_1 (quantity) and y_2 (price), the exogenous variable is x (income), and the disturbances are u_1 (demand shock) and u_2 (supply shock). For convenience the intercepts are suppressed in both equations. The *structural form* of the model is

$$\text{Demand } y_1 = \alpha_1 y_2 + \alpha_2 x + u_1, \tag{1.1}$$

$$\text{Supply } y_2 = \alpha_3 y_1 + u_2. \tag{1.2}$$

With the terms in y_1 and y_2 transferred to the left-hand side, the matrix representation of structural form is

$$(y_1, y_2)\begin{pmatrix} 1 & -\alpha_3 \\ -\alpha_1 & 1 \end{pmatrix} = x(\alpha_2, 0) + (u_1, u_2),$$

or

$$\mathbf{y}'\mathbf{\Gamma} = \mathbf{x}'\mathbf{B} + \mathbf{u}'$$

In the structural-form coefficient matrices $\mathbf{\Gamma}$ and $\mathbf{B}$, the columns refer to equations, while the rows refer to variables.

Each endogenous variable can be solved for in terms of the exogenous variables and structural shocks to get the *reduced form* of the model:

$$\text{Quantity } y_1 = \pi_{11} x + v_1, \tag{1.3}$$

$$\text{Price } y_2 = \pi_{12} x + v_2. \tag{1.4}$$

In matrix form,

$$(y_1, y_2) = x(\pi_{11}, \pi_{12}) + (v_1, v_2),$$

or

$$\mathbf{y}' = \mathbf{x}'\mathbf{B}\mathbf{\Gamma}^{-1} + \mathbf{u}'\mathbf{\Gamma}^{-1} = \mathbf{x}'\mathbf{\Pi} + \mathbf{v}'$$

The reduced form is derived by post-multiplying the structural form by $\mathbf{\Gamma}^{-1}$, where $\mathbf{\Pi} = \mathbf{B}\mathbf{\Gamma}^{-1}$ is the reduced-form coefficient matrix and $\mathbf{v}' = \mathbf{u}'\mathbf{\Gamma}^{-1}$ is the reduced-form disturbance vector.

Next we consider the statistical specification of a linear simultaneous-equation model for the general case of a $m \times 1$ endogenous-variable vector $\mathbf{y}$, the $k \times 1$ exogenous-variable vector $\mathbf{x}$ and the $m \times 1$ structural-disturbance vector $\mathbf{u}$. The specification is the following:

$$\mathbf{y}'\mathbf{\Gamma} = \mathbf{x}'\mathbf{B} + \mathbf{u}', \tag{A1}$$

$$\mathbf{\Gamma} \text{nonsingular}, \tag{A2}$$

$$E(\mathbf{u}|\,\mathbf{x}) = 0, \tag{A3}$$

$$V(\mathbf{u}|\,\mathbf{x}) = \mathbf{\Sigma} \text{ positive definite}. \tag{A4}$$

Here $\mathbf{\Gamma}$ is $m \times m$, $\mathbf{B}$ is $k \times m$, $\mathbf{\Sigma}$ is $m \times m$. Assumption (A1) gives the system of m structural equations in m endogenous variables. Assumption (A2) says that the system is complete in the sense that $\mathbf{y}$ is uniquely determined by $\mathbf{x}$ and $\mathbf{u}$. (A3) says that $\mathbf{x}$ is exogenous in the sense that the conditional expectation of $\mathbf{u}$ given $\mathbf{x}$ is zero for all values of $\mathbf{x}$. Assumption (A4) is a homoskedasticity requirement, and positive definiteness rules out exact linear dependency among the structural disturbances.

The implications of the specification (A1)–(A4) are the following:

$$\mathbf{y}' = \mathbf{x}'\mathbf{\Pi} + \mathbf{v}', \mathbf{v}' = \mathbf{u}'\mathbf{\Gamma}^{-1}, \tag{B1}$$

$$E(\mathbf{v}|\,\mathbf{x}) = 0, \tag{B2}$$

$$V(\mathbf{v}|\,\mathbf{x}) = \left(\mathbf{\Gamma}^{-1}\right)\mathbf{\Sigma}\mathbf{\Gamma}^{-1} = \mathbf{\Omega} \text{ positive definite.} \quad \text{(B3)}$$

The reduced-form disturbance vector $\mathbf{v}$ is mean-independent of, and homoskedastic with respect to, the exogenous variable vector $\mathbf{x}$.

Next we turn from the population to the sample. We suppose that a sample of n observations from the multivariate distribution of $\mathbf{x}$ and $\mathbf{y}$ is obtained by stratified sampling: n values of $\mathbf{x}$ are selected, forming the rows of the $n \times k$ observed matrix $\mathbf{X}$ with rank $(\mathbf{X}) = k$. For each observation, a random drawing is made from the relevant conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, giving the rows of the $n \times m$ observed matrix $\mathbf{Y}$, where the successive drawings are independent. The statements about asymptotic properties of the estimators rely on the additional assumption that the matrix $\mathbf{X}'\mathbf{X}/n$ has a positive definite limit. If instead sampling is random from the joint distribution of $\mathbf{x}$ and $\mathbf{y}$, there is no substantial change in the results.

Ordinary Least Squares

In simultaneous equations models, the parameters of interest are the structural parameter, the $\alpha's$ in the demand–supply example and the elements of $\mathbf{\Gamma}$ and $\mathbf{B}$ and in general case, rather than the reduced form parameters, the $\pi's$ or $\mathbf{\Pi}$. Ordinary least squares (OLS) estimation of the structural parameters is not appropriate because the structural parameters are not coefficients of the conditional expectation functions among the observable variables. We now illustrate this point for the supply equation of the demand and supply model.

The reduced form of the demand and supply model expressed explicitly in terms of the structural parameters:

$$y_1 = (\alpha_2 x + \alpha_1 u_2 + u_1)/(1 - \alpha_1\alpha_3), \quad (2.1)$$

$$y_2 = (\alpha_2\alpha_3 x + \alpha_3 u_1 + u_2)/(1 - \alpha_1\alpha_3). \quad (2.2)$$

For convenience, suppose that x, u_1 and u_2 are trivariate-normally distributed with zero means, variances $\sigma_x^2, \sigma_1^2, \sigma_2^2$, and zero correlations. Then y_2 and y_1 are bivariate normal, so the conditional expectation of y_2 given y_1 is

$$E(y_2|\,y_1) = \alpha^* y_1$$

with

$$\alpha^* = C(y_1, y_2)/V(y_1).$$

If $a^* = \alpha_3$, then the sample least squares regression of y_2 on $y_1$will provide a unbiased minimum variance estimator of α_3. If, $a^* \neq \alpha_3$, then least squares is not appropriate for the estimation of α_3.

From Eqs. (2.1) and (2.2) we calculate

$$C(y_1, y_2) = \left(\alpha_2\alpha_2\sigma_x^2 + \alpha_3\sigma_1^2 + \alpha_1\sigma_2^2\right) \times/(1 - \alpha_1\alpha_3)^2,$$

$$V(y_1) = \left(\alpha_2^2\sigma_x^2 + \sigma_1^2 + \alpha_1^2\sigma_2^2\right)/(1 - \alpha_1\alpha_3)^2.$$

Let $\theta = \alpha_2^2\sigma_x^2 + \sigma_1^2$. Then

$$\alpha^* = \alpha_3 \frac{\theta}{\left(\theta + \alpha_1^2\sigma_2^2\right)} + \frac{\alpha_1\sigma_2^2}{\theta + \alpha_1^2\sigma_2^2}.$$

Clearly the parameter of interest α_3 is not the slope of the conditional expectation function of y_2 given y_1. This result is usually described by saying that OLS gives a biased estimator of the structural parameter α_3. Another description is that OLS gives a unbiased estimator of slope of the conditional expectation function, which happens to differ from the slope of the structural equation. Observe that $a^* = \alpha_3$ in the special case with $\alpha_1 = 0$; in this case, y_1 is a function of x and u_1 only so that $E(u_2|\,y_1) = 0$.

The problem with OLS can be illustrated without relying on normality. From (1.2) we get

$$E(y_2|\,y_1) = \alpha_3 y_1 + E(u_2|\,y_1).$$

From Eq. (2.2),

$$C(y_1, u_2) = C(\alpha_2 x + \alpha_1 u_2 + u_1, u_2)/(1 - \alpha_1\alpha_3) = \alpha_1\sigma_2^2/(1 - \alpha_1\alpha_3).$$

Because y_1 and u_2 are correlated, we see that $E(u_2|\,y_1) \neq E(u_2) = 0$.

Indirect Least Squares

The next method we consider uses OLS to estimate the reduced-form parameters, and then converts the OLS reduced-form estimates into estimates of the structural form parameters. This method, called 'indirect least squares' (ILS), produces estimates that are consistent, although not unbiased. Koopmans and Hood (1953) attribute ILS to M.A. Girshick. Again see Farebrother (1999) for precursors.

The key to ILS is the relation that relates the reduced-form coefficients to the structural-form coefficients, namely, $\mathbf{\Pi} = \mathbf{B}\mathbf{\Gamma}^{-1}$, which can be rewritten as $\mathbf{\Pi}\mathbf{\Gamma} = \mathbf{B}$. Suppose $\mathbf{\Pi}$ is known along with the prior knowledge that certain elements of $\mathbf{\Gamma}$ and $\mathbf{B}$ are zero. The question is whether we can solve $\mathbf{\Pi}\mathbf{\Gamma} = \mathbf{B}$ uniquely for the remaining unknown elements of $\mathbf{\Gamma}$ and $\mathbf{B}$. When a structural parameter is uniquely determined, we say that the parameter is *identified in terms of* $\mathbf{\Pi}$ or, more simply, that is identified. This suggests that the identified structural-form parameters may be estimated via OLS estimates of the reduced-form coefficients.

The relation between reduced-form and structural coefficients for the demand and supply model is the following:

$$(\pi_{11} \quad \pi_{12}) \begin{pmatrix} 1 & -\alpha_3 \\ -\alpha_1 & 1 \end{pmatrix} = (\alpha_2 \quad 0).$$

There are two equations in three unknowns:

$$\pi_{11} - \alpha_1 \pi_{12} = \alpha_2 \pi_{12} - \alpha_3 \pi_{11} = 0. \tag{3.1}$$

On the right-hand-side of (3.1), solve the equation for $\alpha_3 = \pi_{12}/\pi_{11}$. We conclude that the slope coefficient of the supply equation is identified. With respect to estimation, the ILS estimate of α_3 is obtained by replacing π_{11} and π_{12} by their OLS counterparts.

The ILS estimator of α_3 is consistent since the equation-by-equation OLS estimators of π_{11} and π_{12} are consistent. Moreover, the equation-by-equation OLS estimates are the same as the generalized least squares (GLS) estimates, that is, the OLS and GLS estimates coincide in every sample. This is because the explanatory variables are identical in the two reduced-form eqs. A consequence is that the ILS estimator is asymptotically efficient.

Indirect Feasible Generalized Least Squares

For some simultaneous-equation models, prior knowledge that certain elements of $\mathbf{\Gamma}$ and $\mathbf{B}$ are zero implies restrictions on $\mathbf{\Pi}$. In this situation, equation-by-equation OLS estimates of the π's are not optimal, and ILS does not yield a unique estimate of the structural parameters. We now illustrate the case with restrictions on $\mathbf{\Pi}$ using a modification of the original structural model.

The modified model has three exogenous variables, x_1 (income), x_2 (wage rate) and x_3 (interest rate). The modification consists of allowing the three exogenous variables to enter the supply equation:

$$\text{Demand } y_1 = \alpha_2 y_2 + \alpha_1 x_1 + u_1, \tag{4.1}$$

$$\text{Supply } y_2 = \alpha_3 y_1 + \alpha_4 x_1 + \alpha_5 x_2 + \alpha_6 x_3 + u_2. \tag{4.2}$$

The reduced-form of the modified structural-form system is

$$\text{Quantity } y_1 = \pi_{11} x_1 + \pi_{21} x_2 + \pi_{31} x_3 + v_1, \tag{4.3}$$

$$\text{Price } y_2 = \pi_{12} x_1 + \pi_{22} x_2 + \pi_{32} x_3 + v_2 \tag{4.4}$$

In the $\mathbf{\Pi}\mathbf{\Gamma} = \mathbf{B}$ format, the relation between the reduced-form and structural coefficients is:

$$\begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \end{pmatrix} \begin{pmatrix} 1 & -\alpha_3 \\ -\alpha_1 & 1 \end{pmatrix} = \begin{pmatrix} \alpha_2 & \alpha_4 \\ 0 & \alpha_5 \\ 0 & \alpha_6 \end{pmatrix}.$$

There are now six equations in six unknowns:

$$\pi_{11} - \alpha_1 \pi_{12} = \alpha_2 \quad \pi_{12} - \alpha_3 \pi_{11} = \alpha_4$$

$$\pi_{21} - \alpha_1 \pi_{22} = 0 \quad \pi_{22} - \alpha_3 \pi_{21} = \alpha_5 \tag{4.5}$$

T

$$\pi_{31} - \alpha_1\pi_{32} = 0 \quad \pi_{32} - \alpha_3\pi_{31} = \alpha_6.$$

The system on the left of (4.5) determines the parameters of the demand equation. Solve either of the equations that has 0 on its right-hand side for $\alpha_1 = \pi_{31}/\pi_{32} = \pi_{21}/\pi_{22}$, and then get α_2 from the remaining equation. Clearly, the coefficients of the demand equation are identified in terms of $\mathbf{\Pi}$. Furthermore, there is a restriction on the π's, namely $\pi_{31}/\pi_{32} = \pi_{21}/\pi_{22}$, because on the left of (4.5) there are three equations in two unknowns.

The system on the right-hand side of (4.5), which refers to the supply equation, consists of three equations in four unknowns. Once a value is assigned to α_3, the equations can be solved for α_4, α_5, α_6. A different arbitrary value for α_3 generates different values for α_4, α_5, α_6. The solution is not unique. Hence, the coefficients of the supply equation are not identified in terms of $\mathbf{\Pi}$.

With respect to estimation, ILS using the equation-by-equation OLS estimates of $\mathbf{\Pi}$ will not give unique estimates of the structural parameters of the supply equation. The result is two different ILS estimates of α_1. This problem can be overcome by estimating the reduced-form subject to the restriction $\pi_{31}/\pi_{32} = \pi_{21}/\pi_{22}$. The restricted estimates of the $\pi's$ can be converted into unique estimates of the α's using the sample counterpart of the system (4.5).

Suppose there are restrictions on $\mathbf{\Pi}$. Then the fact that the explanatory variables are identical in every reduced-form equation does not imply that the OLS and GLS estimates of the π's are the same. In other words, OLS estimation of the reduced form will not be optimal. If the variance matrix of the reduced-form disturbance vector $\mathbf{\Omega}$ is known, then GLS subject to the restrictions on $\mathbf{\Pi}$ is the natural (nonlinear) estimation procedure. The conversion of the GLS estimates of the π's into estimates of the α's can be described as 'indirect GLS'. Since the GLS estimator is consistent and asymptotically efficient, the indirect-GLS estimators of $\mathbf{\Gamma}$ and $\mathbf{B}$ are also consistent and asymptotically efficient.

When $\mathbf{\Omega}$ is unknown, as is true in practice, feasible GLS is the natural estimation procedure for $\mathbf{\Pi}$. Feasible GLS is similar to GLS except that an estimator $\hat{\mathbf{\Omega}}$ is used in place of $\mathbf{\Omega}$. The estimator $\hat{\mathbf{\Omega}}$ comes from the residuals of the equation by equation OLS reduced-form regressions. The resulting estimates of the α's are referred to as 'indirect-FGLS' estimates because the FGLS estimates of the π's are converted into estimates of the α's. Because the FGLS estimator of $\mathbf{\Pi}$ is consistent and asymptotically efficient, the indirect-FGLS estimators of $\mathbf{\Gamma}$ and $\mathbf{B}$ are also consistent and asymptotically efficient.

Indirect GLS and indirect FGLS are referred to as 'full-information' methods because they use all the restrictions on $\mathbf{\Pi}$ at once. Estimation of a single structural equation using only the restrictions on $\mathbf{\Pi}$ for that equation alone is often called 'limited information' estimation. If all the restrictions are correctly specified, then full-information estimators are more efficient than limited-information estimators.

In some variants of the simultaneous-equation model it is assumed that $\mathbf{u}|\mathbf{x}$ is multivariate normal. The addition of the normality assumption enables the estimation of $\mathbf{\Pi}$ by maximum likelihood. The resulting estimator of the structural parameters is known as 'full-information maximum likelihood', or FIML. If $\mathbf{\Omega}$ is known, then FIML coincides with indirect-GLS. If $\mathbf{\Omega}$ is unknown, FIML differs from indirect-FGLS, but the estimators have the same asymptotic distribution.

The difference between FIML and indirect FGLS can be clarified by briefly turning from the population to the sample. Let, $\hat{\mathbf{V}} = \mathbf{Y} - \mathbf{XP}$, where $\mathbf{P} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the estimator of $\mathbf{\Pi}$ obtained by equation-by-equation OLS. The estimator of $\mathbf{\Omega}$ used in FGLS is $\hat{\mathbf{\Omega}} = (1/n)\hat{\mathbf{V}}\hat{\mathbf{V}}$. The criterion minimized by FGLS is $\mathrm{tr}\left(\hat{\mathbf{\Omega}}^{-1}\mathbf{VV}\right)$, where $\mathbf{V} = \mathbf{Y} - \mathbf{X}\ \mathbf{\Pi}$. FIML proceeds by inserting $\hat{\mathbf{\Omega}} = (1/n)\hat{\mathbf{V}}\hat{\mathbf{V}}$ (as a conditional solution) into the log-likelihood function to obtain the log-likelihood concentrated on $|\mathbf{V}'\mathbf{V}|$. The consequence is that the criterion minimized by FIML is $|\mathbf{V}'\mathbf{V}|$. The difference in the criteria explains the difference in the estimators.

The maximum likelihood estimation of a single structural-form equation that uses only the restrictions on $\mathbf{\Pi}$ for that equation alone is referred to as 'limited-information maximum likelihood', or LIML. We next consider another limited-information estimation method.

Two-Stage Least Squares

The 2SLS estimator uses the unrestricted reduced-form estimate **P**, the equation-by-equation OLS estimates of the π's, which accounts for its popularity. The mechanics of the 2SLS method can be described simply. In the first stage, the right-hand-side endogenous variables of the structural equation are regressed on all the exogenous variables in the reduced form, and the fitted values are obtained. In the second stage, the right-hand-side endogenous variables are replaced by their fitted values, and the left-hand-side endogenous variable of the equation is regressed on the right-hand- side fitted values and the exogenous variables included in the equation.

Two rationales for the 2SLS procedure are now developed using the demand equation of the modified structural model. The starting point for the first rationale is the expectation of the demand equation conditional on x_1, x_2, and x_3. Taking expectations gives

$$E(y_1|x_1,x_2,x_3) = \alpha_1 E(y_2|x_1,x_2,x_3) + \alpha_2 x_1 + E(u_1|x_1,x_2,x_3),$$

or

$$E(y_1|x_1,x_2,x_3) = \alpha_1 y_2^* + \alpha_2 x_1.$$

From the reduced-form Eq. (4.4), $y_2^* = E(y_2|x_1,x_2,x_3) = \pi_{12}x_1 + \pi_{22}x_2 + \pi_{32}x_3$. Because y_2^* is linear function of the exogenous variables, it is exogenous. If y_2^* were observed, then y_1 could be regressed on y_2^* and x_1 to get unbiased estimates of α_1 and α_2. But y_2^* is unobservable because π_{12}, π_{22} and π_{32} are unknown. However, an unbiased and consistent estimate $\hat{y}_2$ can be obtained by replacing the unknown π's by their OLS estimates. Then, making the replacement of $\hat{y}_2$ for y_2^* produces consistent estimates of the structural parameters.

The second rationale exploits the fact that in the population the following moment conditions hold:

$$E(u_1) = 0, \quad C(x_1,u_1) = C(x_2,u_1) = C(x_3,u_1) = 0.$$

These imply two orthogonality conditions:

$$E(y_2^* u_1) = 0, E(x_1,u_1) = 0.$$

If we let $u_1 = y_1 - (a_1 y_2^* + a_2 x_1)$, then α_1 and α_2 are the values for a_1 and a_2 that make $E(y_2^* u_1) = 0$ and $E(x_1 u_1) = 0$.2SLS chooses the estimates that make the analogous sample quantities zero, that is, $\sum_i \hat{y}_{2i} u_{1i} = 0$ and $\sum_i x_{1i} u_{1i} = 0$, ($i = 1, \ldots, n$). This illustrates that 2SLS has an *instrumental-variable* (IV) interpretation.

The IV interpretation can be illustrated more explicitly by writing the demand equation in terms of the observations for a sample size of n:

$$\mathbf{y}_1 = \mathbf{y}_2\alpha_1 + \mathbf{x}_1\alpha_2 + \mathbf{u}_1 = (\mathbf{y}_2 \ \ \mathbf{x}_1)\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} + \mathbf{u}_1 = \mathbf{Z}_1\boldsymbol{\alpha} + \mathbf{u}_1,$$

where in the context of the demand equation $\mathbf{y_1}$ and $\mathbf{y_2}$ are the columns of $\mathbf{Y}$ and $\mathbf{x_1}$ is the first column of $\mathbf{X}$. As we have shown, regressing $\mathbf{y_1}$ on $\mathbf{Z_1}$ will not give a consistent estimator for $\boldsymbol{\alpha_1}$. Instead replace $\mathbf{Z_1}$ by

$$\hat{\mathbf{Z}}_1 = \mathbf{N}\mathbf{Z}_1 = \mathbf{N}(\mathbf{y}_2,\mathbf{x}_1) = (\mathbf{N}\mathbf{y}_2, \mathbf{N}\mathbf{x}_1) = (\hat{\mathbf{y}}_2, \mathbf{x}_1),$$

where $\mathbf{N}$ is the idempotent matrix $\mathbf{X}(\mathbf{X}'\mathbf{X}^{-1})\mathbf{X}'$. Regressing $\mathbf{y_1}$ on $\hat{\mathbf{Z}}_1$ gives the normal equations,

$$\hat{\mathbf{Z}}_1'\hat{\mathbf{Z}}_1\mathbf{a}_1^* = \hat{\mathbf{Z}}_1'\mathbf{y}_1,$$

the solution to which is the 2SLS estimator.

The 2SLS normal equations are equivalent to a set of orthogonality conditions: $\hat{\mathbf{Z}}_1'\mathbf{u}_1 = 0$, where $\mathbf{u}_1 = \mathbf{y}_1 - \mathbf{Z}_1\mathbf{a}_1^*$. The equivalence follows from an algebraic fact:

$$\hat{\mathbf{Z}}_1'\mathbf{Z}_1 = \mathbf{Z}_1'\mathbf{N}'\mathbf{Z}_1 = \mathbf{Z}_1'\mathbf{N}'\mathbf{N}\mathbf{Z}_1 = \hat{\mathbf{Z}}_1'\hat{\mathbf{Z}}_1.$$

The variables in $\hat{\mathbf{Z}}_1$ are legitimate instruments because they are, at least asymptotically, uncorrelated with the disturbance. The IV interpretation implies that the 2SLS estimator is consistent. In fact, it is the optimal feasible IV estimator.

We also note that the 2SLS estimator can be interpreted as a general-method-of moments (GMM) estimator. In the above example, this follows from the fact that a minimizes the quadratic form

$$\left(\mathbf{y}_1 - \mathbf{Z}_1\mathbf{a}_1^*\right)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{y}_1 - \mathbf{Z}_1\mathbf{a}_1^*\right).$$

It can be shown that 2SLS is the optimal feasible GMM estimator. An advantage of the GMM approach is that heteroskedasticity and autocorrelation can be accommodated by an appropriate redefinition of the optimal weighting matrix in the definition of the GMM estimator (see Ruud 2000, pp. 718–21).

We conclude this section with some remarks on estimation in the simultaneous-equations model.

1. If all the structural equations are identified, and there are no restrictions on $\mathbf{\Pi}$, then ILS, indirect-FGLS, FIML, LIML and 2SLS all produce the same estimates.
2. If there are restrictions on $\mathbf{\Pi}$, then LIML and 2SLS produce different estimates in the sample, but the estimators have the same asymptotic distribution, and similarly for indirect FGLS and FIML.
3. If a parameter is not identified, then there is no method to estimate it consistently.
4. We have confined our attention to the case in which the prior information used for identification consists of normalizations and exclusions (zero restrictions). If other information is available (for example, $\mathbf{\Sigma}$ is diagonal, or a coefficient in one structural equation is equal to a coefficient in another), then some modifications are needed in the description of the estimators and their statistical properties.

The *K-Class* Family

The *k*-class family of estimators of the coefficients of a single structural equation is illustrated for the demand equation of the modified structural model. For this equation, the estimator is

$$\mathbf{a}_k^* = (\mathbf{Z}_1'(\mathbf{I} - k\mathbf{M})\mathbf{Z}_1{}^{-1}\mathbf{Z}_1'(\mathbf{I} - k\mathbf{M})\mathbf{y}_1,$$

where $\mathbf{M} = \mathbf{I} - \mathbf{N}$. This family was introduced by Theil (1953b, 1961). It includes the OLS estimator ($k = 0$) and the 2SLS estimator ($k = 1$).

A remarkable fact is that the *k*-class family includes LIML. The LIML estimator is obtained by setting $k = \lambda^*$, where λ^* the smallest root of

$$|\mathbf{W}_1 - \lambda\mathbf{W}| = 0.$$

In the determinantal equation, $\mathbf{W_1}$ is the cross-product matrix of residuals from the OLS regression of $(\mathbf{y_1}, \mathbf{y_2})$ on $\mathbf{x_1}$ (the included endogenous variables on the included exogenous variable), and $\mathbf{W}$ is the cross-product matrix of the residuals from the OLS regression of $(\mathbf{y_1}, \mathbf{y_2})$ on $\mathbf{X}$ (the included endogenous on all the exogenous variables). Moreover, the LIML estimator of α_1 is the value of a_1 that minimizes the variance ratio:

$$\frac{(1, -a_1)\mathbf{W}_1(1, -a_1)'}{(1, -a_1)\mathbf{W}(1, -a_1)'}.$$

Accordingly, the LIML estimator is also sometimes referred to as the 'least-variance ratio' estimator.

The *k*-class estimator is consistent if $(k - 1)$ converges in probability to 0 and has the same limiting distribution as 2SLS if $n^{1/2}(k - 1)$ converges in probability to 0. These conditions are clearly not satisfied when $k = 0$ and hence by OLS. A proof that these conditions are satisfied for LIML is given in Amemiya (1985, pp. 237–8). Zellner (1998) shows that a Bayesian method-of-moments approach justifies certain other members of the *k*-class (and double *k*-class) family of estimators.

Finite Sample Distributions

There has been debate about whether the LIML estimator is better than 2SLS. There a son for this debate is that the finite sample distributions of the estimators differ when there are restrictions on $\mathbf{\Pi}$. Hence we now limit our attention to the case where restrictions are present.

A key difference between the estimators is the existence of moments. The 2SLS estimator has moments up to certain order. By contrast, the LIML estimator has no moments. This result holds for an arbitrary number of included endogenous variables. Mariano (2001) reviews the moment existence results.

In addition to the moment existence results, closed form expressions for the moments and probability densities of 2SLS, LIML and *k*-class estimators have been derived. These expressions are complicated; see Phillips (1983) for specific references.

The finite sample results have mostly come from the study of a model with two included endogenous variables. Moreover, it is often assumed that the disturbances are normal. The finite sample results are obtained using analytical and simulation methods. A survey of the results and their practical implications is given in Mariano (2001). One of the results is that the LIML distribution is far more symmetric than 2SLS, though more spread out, and it approaches normality faster. Similarly, Anderson (1982) concludes that for many cases that occur in practice the standard the normal theory is inadequate for 2SLS, but provides a fairly good approximation to the actual distribution of the LIML estimator. The symmetry result is not surprising because the approximate distribution of the LIML estimator (obtained from large sample asymptotic expansions) is median unbiased. Median unbiasedness does not hold in general for 2SLS.

It is helpful to put the debate over the relative merits of LIML and 2SLS in perspective. On the assumption that the model is correctly specified, the presumption is that maximum likelihood will have better properties in some well-defined sense. This is because it uses more information than GMM. See Anderson et al. (1986) and Takeuchi and Morimume (1985) for results on the second-order efficiency of LIML. An advantage of GMM is that it is robust to misspecifications of the disturbance distribution, although this advantage was not the original motivation for introducing 2SLS.

Epilogue

Keynesian economics initially played a key role in propelling research into the estimation of simultaneous-equations models. Interest in the estimation of linear simultaneous-equations models began to wane from about the late 1970s. Historically, this decline paralleled the decline of the Keynesian paradigm as a result of the monetarist counter-revolution and later rational expectations. At the same time, there was growing awareness that linear simultaneous-equations models often suffered from potentially serious misspecification. On the one hand, even if one takes for granted the statistical specification of the model (A1–A4), economic theory did not provide a satisfactory basis for deciding what endogenous and exogenous variables should be excluded from individual structural equations. As a consequence, the identification of structural parameters was open to question. On the other hand, the statistical specification of the linear structural-form model was itself questionable, due to either the nature of the data or considerations of economic theory or both. The issue of identification was highlighted by Sims (1980) in a well-known paper aptly titled 'Macroeconomics and Reality'.

Simultaneous-equations models have been generalized by the introduction of nonlinear structural equations. Amemiya (1974) generalized the 2SLS method to nonlinear models. His nonlinear 2SLS estimator is a GMM estimator. However, unlike in the case of linear models, it cannot be thought of as being obtained in two steps where the first consists of running a least squares regression. The GMM is a two-step estimator, but the first step consists of choosing a weighting matrix. More generally, GMM and IV estimators can be thought of as the descendants of the 2SLS approach. They constitute the contemporary basis for much of the estimation of structural parameters macroeconomics. It is in this sense that 2SLS lives on as a structural estimation method.

Does the identification and estimation of simultaneous-equations models have a future? There are some positive signs. Although indirectly, these issues continue to play a role in the identified vector autoregression literature, for

example, Bernanke and Blinder (1992), Gordon and Leeper (1995), Cushman and Zha (1997). More recently, there has been a revival of interest in simultaneous-equations models formulated as nonlinear dynamic stochastic general equilibrium models. This formulation has the advantage that the resulting models are viewed as more firmly anchored in economic theory. Estimation of such models presents serious challenges. Some examples where these challenges are addressed include DeJong et al. (2000) and Fernandez-Villaverde and Rubio-Ramirez (2005).

See Also

- ▶ Seemingly Unrelated Regressions
- ▶ Simultaneous Equations Models

Bibliography

Amemiya, T. 1974. The nonlinear two-stage least-stage estimator. *Journal of Econometrics* 2: 105–110.

Amemiya, T. 1985. *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Anderson, T. 1982. Some recent developments on the distribution of single-equation estimators. In *Advances in econometrics*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.

Anderson, T. 2005. Origins of the limited information maximum likelihood and two stage least squares estimators. *Journal of Econometrics* 127: 1–16.

Anderson, T., and H. Rubin. 1949. Estimator of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20: 46–63.

Anderson, T., and H. Rubin. 1950. The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 21: 570–582.

Anderson, T., N. Kunitomo, and K. Morimune. 1986. Comparing single equation estimators in a simultaneous equation system. *Econometric Theory* 2: 1–32.

Basmann, R. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25: 77–83.

Bernanke, B., and A. Blinder. 1992. The federal funds rate and the channels of monetary transmission. *American Economic Review* 82: 901–921.

Cushman, D., and T. Zha. 1997. Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary Economics* 39: 433–448.

DeJong, D., B. Ingram, and C. Whiteman. 2000. A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* 98: 203–223.

Farebrother, R. 1999. *Fitting linear relationships: A history of the calculus of observations 1750–1900*. New York: Springer.

Fernandez-Villaverde, J., and J. Rubio-Ramirez. 2005. Estimating dynamic equilibrium economies: Linear versus nonlinear likelihood. *Journal of Applied Econometrics* 20: 891–910.

Goldberger, A. 1991. *A course in econometrics*. Cambridge, MA: Harvard University Press.

Gordon, D., and E. Leeper. 1995. The dynamic impacts of monetary policy: An exercise in tentative identification. *Journal of Political Economy* 102: 1228–1247.

Hood, W., and T. Koopmans. 1953. *Studies in econometric method*, Cowles foundation monograph 14. New Haven: Yale University Press.

Koopmans, T. 1950. *Statistical inference in dynamic economic models*, Cowles commission monograph 10. New York: John Wiley and Sons.

Koopmans, T., and W. Hood. 1953. The estimation of simultaneous linear economic relationships. In *Studies in econometric method*, Cowles foundation monograph 14, ed. W. Hood and T. Koopmans. New Haven: Yale University Press.

Mariano, R. 2001. Simultaneous equation model estimators: Statistical properties and practical implications. In *A companion to theoretical econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Mittelhammer, R., G. Judge, and D.J. Miller. 2000. *Econometric foundations*. Cambridge: Cambridge University Press.

Phillips, P. 1983. Exact small sample theory in the simultaneous equation model. In *Handbook of econometrics*, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Ruud, P. 2000. *An introduction to classical econometric theory*. Oxford: Oxford University Press.

Sargan, J. 1958. Estimation of economic relationships using instrumental variables. *Econometrica* 67: 557–586.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.

Takeuchi, K., and K. Morimune. 1985. Third-order efficiency of the extended maximum likelihood estimator in a simultaneous equation system. *Econometrica* 53: 177–200.

Theil, H. 1953a. Repeated least-squares applied to a complete equation systems. Mimeo. The Hague: Central Planning Bureau.

Theil, H. 1953b. Estimation and simultaneous correlation in complete equation systems. Mimeo. The Hague: Central Planning Bureau.

Theil, H. 1961. *Economic forecasts and policy*. 2nd ed. Amsterdam: North- Holland.

Zellner, A. 1998. The finite sample properties of simultaneous equations' estimates and estimators: Bayesian and non-Bayesian approaches. *Journal of Econometrics* 83: 185–212.